THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *27 Novembre 2014* **par :**

MONIA BEN MLOUKA

Le référencement en Langue des signes : analyse et reconnaissance du pointé.

JURY

MME ANNELIES BRAFFORT	Directrice de recherche au CNRS, LIMSI-ORSAY, Paris	Rapporteur
MME AGNÈS MILLET	Professeur émérite d'université, Grenoble	Rapporteur
M. PHILIPPE JOLY	Professeur d'université, IRIT, Toulouse	Examinateur
M. MICHEL DAYDÉ	Professeur d'université, IRIT, Toulouse	Directeur de thèse

---

École doctorale et spécialité :
MITT : Image, Information, Hypermedia

Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)

Directeur de Thèse :
M. Patrice Dalle

Rapporteurs :
Mme Annelies Braffort et *Mme Agnès Millet*

Remerciements

Ce travail a été réalisé au sein de l'équipe **T**raitement et **C**ompréhension d'**I**mages à l'IRIT (Toulouse) et financé par le projet *Sign'Com*.

Je tiens à exprimer ma profonde reconnaissance et mon respect à M. Patrice Dalle qui m'a fait découvrir l'univers fascinant de la langue des signes. Sa disparition au mois d'Août dernier était une profonde douleur. Je le remercie pour le temps qu'il m'a consacrée durant ces années de travail, des heures de discussions sur les énoncés signés, du plus simple au plus complexe. M. Patrice Dalle était également mon enseignant en matière de précision, de critique et surtout d'auto-critique. L'aventure a commencé par un concours de circonstances lié au thème du regard auquel je me suis un peu trop attachée à tel point que j'ai gardé ce thème dans mon titre de sujet de thèse même si le sujet a considérablement évolué. En passant par l'espace de signation, les locus et au final les gestes. Finalement, le titre a été synchronisé avec l'état actuel de mon travail...mais le thème du regard reste gravé dans ma mémoire.

J'adresse mes sincères remerciements aux membres du jury : M. Michel Daydé en tant que directeur de thèse et M. Philippe Joly en tant qu'examinateur. Je remercie Mme Annalies Braffort et Mme Agnès Millet d'avoir accepté de relire mon manuscrit de thèse et de faire partie de mon jury de thèse. Je tiens, aussi, à exprimer ma sincère reconnaissance à M. Braffort, ses précieux conseils, son écoute et sa disponibilité durant les dernières modifications du manuscrit.

Ce travail a été enrichi par la collaboration avec plusieurs équipes de recherches. L'équipe partenaire du projet Sign'Com, l'IRISA. L'équipe partenaire du projet Dicta-Sign, M&TALS. L'équipe de recherche en linguistique de la langue des signes Québécoise, Marqspat. Je remercie L'entreprise Websourd de nous avoir offert les meilleures conditions pour réaliser le dernier corpus LSF dans leurs locaux.

Je tiens également à remercier Mme Martine Labruyère pour son soutien moral tout au long de ces deux dernières années.

Je veux adresser ma vive reconnaissance à ma collègue de bureau, Juliette Dalle qui m'a beaucoup aidée pour l'élaboration de corpus, avec qui j'ai eu des échanges marquants autour de la structure de la langue des signes française. En réalité, cette thèse m'a permis aussi de rencontrer des linguistes avec qui j'ai eu des échanges, qui étaient pour moi, une découverte à chaque fois. Isabelle Estève, Leila Boutora et Manel Khayech Chaibi à qui j'adresse mes sincères remerciements.

Je remercie mon ancien collègue, François Lefebvre-Albaret pour le temps qu'il a consacré, à la débutante que j'étais, pour m'initier aux notions de bases de la langue des signes française et pour avoir revu mes premières productions scientifiques. Je remercie Romain qui m'a beaucoup aidée dans l'utilisation de la bibliothèque OpenCV.

Je remercie Christophe Collet qui a eu la gentillesse de m'accompagner dans les

dernières étapes de mon travail, Alain Crouzil pour sa réactivité et son aide sur les plans moral et administratif, Denis Kouamé et Adrian Basarab pour leurs précieux conseils.

Je remercie Arturo pour ses encouragements, sa disponibilité et sa gentillesse. Je remercie Philippe, Anthony, Mathieu Muratet, Samir, Laure, Maroua, Bochra, Patrice et Ezgei d'avoir amené le sourire et la bonne ambiance à l'IRIT malgré les moments difficiles de la vie d'un doctorant.

Je remercie ma famille de me soutenir et d'avoir supporté ces longues distances de séparation afin que je puisse réaliser mon rêve. Les mots ne suffiront pas pour exprimer ma reconnaissance à Mathieu G., qui m'apporte un soutien inconditionnel depuis ces deux dernières années, sans oublier son aide en Latex.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction : Les langues des signes

Ce chapitre a pour rôle d'introduire la problématique de cette thèse qui porte sur les langues des signes et leur traitement automatique.

Sommaire

1.1 Les langues des signes

Les langues des signes (LS) sont les langues naturelles utilisées par les sourds et certains entendants côtoyant des locuteurs sourds. Ces langues, dites visuo-gestuelles (car émises par le corps et reçues via la vision), s'expriment dans l'espace en face du locuteur, au moyen d'unités gestuelles composées de gestes des mains et des bras, de mouvements du buste, des épaules et de la tête, d'expressions du visage et de directions du regard, réalisés simultanément. De même que pour les langues vocales, il n'existe pas de LS universelle, mais autant que de communautés différentes de sourds, chaque LS ayant son histoire, ses unités signifiantes et son lexique. Les LS s'inscrivent dans l'espace et dans le temps par des gestes, des mouvements du buste, des épaules et de la tête, des expressions du visage et des jeux de regards, tous signifiants et potentiellement simultanés. Le mode d'expression de ces langues est donc multilinéaire et spatio-temporel. Dans la suite du mémoire, nous utiliserons les termes suivants : paramètre pour désigner les constituants des unités gestuelles, qu'ils soient manuels (configuration, orientation, emplacement et mouvement de la main), ou non manuels (mouvement du buste, des épaules de la tête, des éléments mobiles du visage, ou encore la direction du regard) ; espace de signation, pour désigner l'espace placé devant le signeur et dans lequel s'articule les gestes manuels ; multilinéaire, multilinéarité, pour faire référence au fait que plusieurs paramètres, manuels ou non manuels, peuvent être mis en jeu dans une construction linguistique.

L'espace de signation est composé de zones spatiales qu'on appellera *« référence »* [1] ou *« locus »* [2]. Nous présenterons dans le chapitre suivant deux fonctions linguistiques liées à l'espace de signation ; la construction et l'activation de référence.

Nous utiliserons la notion de signe selon la manière dont il a été réalisé :

- le signe standard qui représente une unité gestuelle qui a un rôle lexical. Les signes standards peuvent être répertoriés dans un dictionnaire de LS ;
- le signe spatialisé ou localisé qui signifie qu'un signe a été associé à un locus ;
- le signe en mouvement dont l'interprétation usuelle est « le verbe directionnel». Nous avons choisi d'employer le terme de signe en mouvement afin d'homogénéiser les termes liés à la notion de signe. Le signe en mouvement se compose d'un emplacement de départ, un mouvement dirigé vers une destination. Exemple les signe [VENDRE](1.6-a) ou [ACHETER](1.6-b) désignant une zone spatiale lié à celui qui réalise l'action (ex. vendeur) et celui qui subit l'action (l'acheteur) ou l'inverse.

Nous utiliserons également les unités gestuelles qui, selon le contexte, seront interprétées différemment (ex. expression faciale particulière, les unités gestuelles de pointage). Par exemple une expression faciale accompagnant un signe. L'exemple (1.2) illustre [AVION DETOURNE] par un déploiement simultané du signe [AVION], d'une rotation de la main et de la tête, une expression faciale particulière ; un froncement des sourcils, un regard vif, exprimant la colère.

FIGURE *1.1 – Signe [AVION]*

Les LS sont des langues naturelles, possédant un lexique et une grammaire, qui ont la particularité d'exploiter de manière pertinente l'iconicité à tous les niveaux linguistiques. Selon les théories linguistiques, cette iconicité peut avoir un statut différent. En France, C. Cuxac a proposé une théorie dans laquelle l'iconicité joue un rôle central [Cuxac 2000]. Il décrit des constructions linguistiques *« illustratives »* qu'il nomme transfert et qui permettent de « dire tout en montrant » (voir les figures 1.3 et 1.4). Il les distinguent

1. la référence est un terme lié à deux fonctions linguistiques construction et activation de référence
2. locus est un terme fréquemment employé par les linguistes. Le pluriel de locus est loci

FIGURE *1.2 – Signe [AVION], une posture et une expression faciale signifiant un avion détourné*

des constructions *« non illustratives »*, qui *« disent sans monter »*, telles que les signes standards (figure 1.1). Le transfert situationnel est une structure de grande conicité qui montre une proximité spatiale entre deux entités spatialisées. L'exemple (1.5 a-b) exprime une proximité spatiale entre un [AVION] et [SOL] puis l'avion posé au sol [3]. Dans cet exemple, la main gauche du signeur, étant immobile, représente le *« localif spatial »* c'est-à dire le le repère sur lequel se base l'identification d'un transfert situationnel.

FIGURE *1.3 – Signe [AVION], un regard qui fixe le signe déployé, signifiant [CET AVION]*

1.2 Le traitement automatique des langues des signes (TALS)

Les études sur les LS sont assez récentes et les connaissances sur lesquelles il est possible de s'appuyer sont peu nombreuses (quand elles ne sont pas soumises à débat).

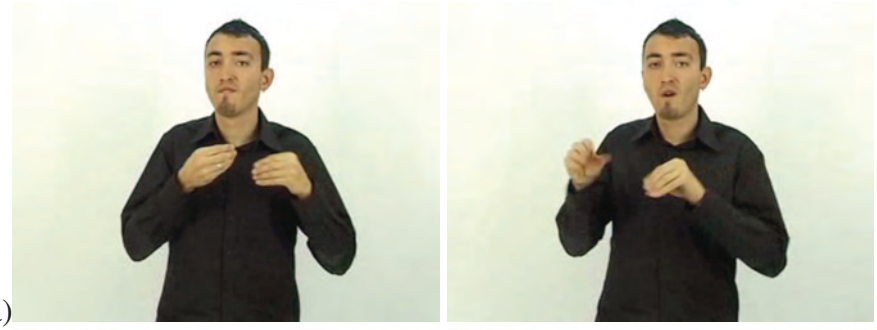
3. L'expression faciale et l'orientation de la main expriment le fait que l'avion s'est écrasé au sol

FIGURE *1.4 – Signe [AVION], un pointage manuel qui désigne le signe déployé, signifiant [CET AVION]*

a)  b) 

FIGURE *1.5 – a) Transfert situationnel exprimé par une proximité spatiale entre les emplacements du signe [AVION] et le signe [SOL] et b) le avion posé sur le sol*

a)



b)

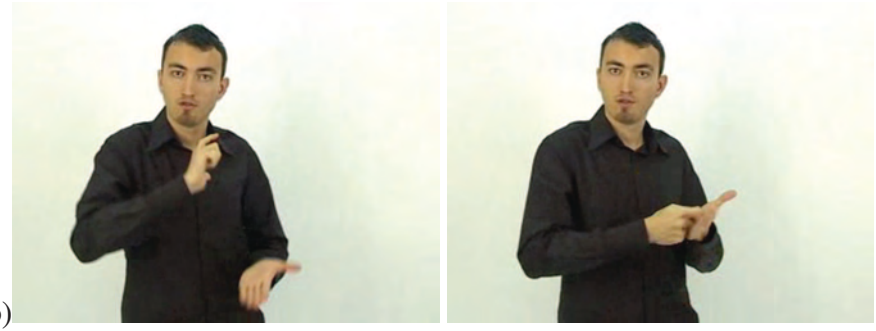


FIGURE *1.6 – Exemples de signes en mouvement : a)le signe [VENDRE] et b) le signe [ACHETER]*

Les LS ne disposent pas de système d'écriture, ce sont des langues de l'oralité : les études s'appuient donc sur des productions « orales » de LS, enregistrées sous forme de corpus vidéo, ou de capture de mouvement depuis peu.

Plusieurs travaux portent sur le traitement automatique de discours en langue des signes. Ces travaux portent sur deux axes principaux ;

- La reconnaissance automatique dans un flux vidéo.
- La synthèse par l'animation de seigneurs virtuels ou avatars signants.

Le cadre de cette thèse s'inscrit dans le premier axe ; plus spécifiquement, dans la reconnaissance automatique de certaines constructions linguistiques en langue des signes. Ce sujet en tant que tel n'a pas été abordé à ce jour. Cependant, quelques travaux, qui ont porté sur l'analyse de caractéristiques de paramètres manuels et non manuels se recoupent avec notre sujet d'étude.

En Langue des signes française, [Lefebvre-Albaret 2010] s'est focalisé sur la caractérisation de la dynamique des gestes afin de repérer automatiquement le début et la fin d'un signe dans un discours signé. [Gonzalez-Preciado 2012] a introduit certains paramètres tel que la configuration manuelle afin d'améliorer le résultat de reconnaissance de signes.

1.3 Contexte de la thèse

Comme nous l'avons mentionné précédemment, le présent travail s'inscrit dans le cadre de l'analyse et de la reconnaissance de constructions linguistiques dans un discours en langue des signes. Plus particulièrement, le travail porte sur l'étude du *référencement*.

Point de vue réalisation gestuelle, le référencement regroupe tous les gestes manuels et non manuels qui permettent de faire référence ou de mettre le focus sur une zone de l'espace de signation. Nous nous sommes intéressés au référencement car il se réalise fréquemment et d'une manière, dans la plus part des cas, multi-linéaire.

Le référencement fait partie de plusieurs constructions linguistiques qui font intervenir les mains, la tête, le tronc ainsi que le regard. Il est présent dans certaines structures de grande conicité telle que le transfert situationnel (Voir exemple 1.5). Dans la réalisation d'un signe en mouvement, le signeur fait référence à la zone de départ du signe puis à celle de destination.

Les linguistes ont observé que les gestes réalisant le référencement sont variés : 1) le mouvement de l'avant bras avec une configuration manuelle précise (main plate ou index tendu), 2) La rotation ou l'inclinaison de la tête, 3) l'orientation du regard, 4) le balancement du tronc et / ou des épaules, 5) la combinaison de certains de ces mouvements.

1.4 Présentation du travail

Dans la section précédente, nous avons montré par des exemples que dans un discours en langue des signes, le référencement se réalise fréquemment faisant intervenir un ou plusieurs gestes déployés de plusieurs manières (variation des unités gestuelles ; la direction, l'emplacement, etc.). Notre objectif est de construire un système de reconnaissance automatique de la construction linguistique de référencement qui tient compte de ces formes. Dans un premier temps, nous nous proposerons d'analyser finement la fréquence et les formes gestuelles du référencement pour les intégrer par la suite dans un système de reconnaissance automatique.

Comme première étape d'analyse, nous étudierons les modèles gestuels de référencement tels qu'ils ont été établis par les linguistes suite à des observations de corpus de langues des signes. En se basant sur les modèles de construction linguistique de référencement, nous identifierons les paramètres à prendre en compte pour construire des modèles informatiques de référencement. L'étape suivante aura comme objectif d'affiner les modèles informatiques construits. Pour cela, nous ferons appel à des méthodes déjà utilisés dans les travaux de TALS, notamment, ceux de [Lefebvre-Albaret 2010] pour l'étude de la dynamique des gestes. Nous proposerons deux type de modèles, temporel et spatial. Ceci veut dire que nous projetterons de quantifier des propriétés :

- temporelles telles que la durée d'un geste et le décalage entre plusieurs gestes dans le cas de référencement multi-linéaire ;
- spatiales telles que la position de la main, l'orientation de la tête et les distances relatives entre la main et la tête quand il s'agit de référencement multi-linéaire.

Cette étape requiert l'utilisation de corpus en langue des signes. Comme étape intermédiaire, nous détaillerons les spécifications des corpus qui seront utilisés dans la construction de modèles. Ce cahier de charge sera également déterminant pour la réutilisation ou pas de certains corpus élaborés dans le cadre de projets antérieurs. Au cours de cette étape intermédiaire, nous utiliserons un protocole de pré-traitement pour faciliter la synchronisation des données du corpus. Nous présenterons dans le troisième chapitre notre choix de corpus de construction de modèles qui se compose de données variées : 1) vidéos annotées, 2) les données de capture de mouvement et 3) les données de capture du regard. Ce choix a été fait suite à des séries d'observations de corpus composés uniquement de vidéos. Nous avons conclu suite à ces observations qu'il serait nécessaire d'enrichir les corpus vidéo avec des données mesurables afin de rendre compte, non pas de la réalisation perçue, mais de la réalisation réelle d'un référencement. Le troisième chapitre présente également les corpus composés uniquement de vidéos annotées utilisés dans les étapes d'exploitation des modèles informatiques $3D$ construits.

La première étape d'exploitation de modèles informatiques de référencement consiste à adapter les modèles $3D$ à un système de reconnaissance $2D$ car l'objectif final étant de détecter automatiquement un référencement dans une vidéo. Ensuite vient l'étape

d'apprentissage qui consiste à établir des seuils pour les paramètres retenus dans les modèles $2D$. Nous décrirons par la suite l'algorithme de détection implémenté. Nous conclurons la description de la méthodologie de reconnaissance par la présentation de quelques résultats de test des modèles $2D$, de leurs interprétations. Nous avons fait le choix de consacrer un chapitre à l'évaluation de l'algorithme implémenté et de la méthodologie de reconnaissance.

Nous récapitulons les étapes de la méthodologie comme suit :

- Chapitre 3 : description des corpus utilisés ;
- Chapitre 4 : construction de modèles informatiques du référencement ;
- Chapitre 5 : présentation de la méthode de passage $3D \rightarrow 2D$, du seuillage des paramètres retenus et de l'algorithme implémenté ;
- Chapitre 6 : évaluation des résultats et de l'algorithme de détection ainsi que de quelques étapes de la méthodologie de reconnaissance.

Avant d'entrer dans les détails de la méthodologie de reconnaissance, nous consacrerons le chapitre suivant à l'étude des modèles linguistiques liés au référencement.

CHAPITRE 2

Le référencement en Langue des Signes

Ce chapitre propose une définition du référencement en langue des signes basée sur des constructions linguistiques faisant intervenir les gestes du corps et l'espace de signation.

Sommaire

2.1 Introduction

Dans cette thèse, nous appelons *référencement* l'action de mentionner quelque chose que l'on connaît au préalable. Par analogie, en langue des signes, nous appelons référencement le fait de concentrer son attention (celle du signeur), pendant un laps de temps, sur une entité du discours. Plus précisément, le référencement pour nous, représente toutes les formes gestuelles qui permettent d'activer une zone spatiale en relation avec le discours signé.

Le référencement peut avoir plusieurs interprétation : 1) un référencement vers une entité spatiale, 2) un référencement vers un signe localisé, 3) un référencement d'une zone de départ et d'une zone d'arrivée (comme c'est le cas du signe [DONNER] - voir la figure 2.1). Ces interprétations font partie de plusieurs modèles linguistiques de la langue des signes tels que les structures de grande iconicité, exemple le transfert situationnel. Dans ce chapitre nous passerons en revue les modèles linguistiques dont les formes gestuelles intègrent la notion de référencement. Il en découlera une catégorisation des formes gestuelles de référencement. Par la suite, nous introduirons la notion de l'espace de signation comme support syntaxique au référencement.

2.2 Les modèles linguistiques liés au référencement

Nous avons constaté que certaines constructions linguistiques font intervenir à la fois le regard, l'espace de signation et les gestes du corps. La structure linguistique de création de référence en fait partie. Dans la section suivante, nous allons passer en revue les modèles linguistiques de construction de référence telles que les structures de grande iconicité de C. Cuxac.

2.2.1 Modèles de création de référence

La création de référence consiste à activer une zone de l'espace, lui associer un signe et y faire référence plus tard dans le discours.

2.2.1.1 Transfert de situation

Le transfert de situation (ou situationnel) fait intervenir le regard. Son rôle est fondamental car il installe le geste effectué par les mains ou la main dominante dans une zone déterminé dans l'espace tri-dimensionnel qui se situe devant le signeur. La fonction de transfert de situation décrit la localisation d'une entité dans l'espace par rapport à la position d'une autre entité associée préalablement à une zone de l'espace. Le regard porte d'abord sur le locatif stable (ou main dominée immobile), puis fixe la main dominante, ensuite, anticipe sur le point d'arrivée de la main dominante.

2.2.1.2 *Transfert de forme*

C'est une fonction qui décrit la construction de la forme d'une entité (exemple [BALLON]) utilisant à la fois la configuration manuelle, le regard et l'espace. Le regard fixe un point de l'espace. Ensuite, un mouvement manuel est effectué pour construire la forme du signe en question. Une phase de stabilisation de la forme dure une à deux images [1] et se situe dans la portion de l'espace fixé par le regard. [Cuxac 2003] mentionne que le regard joue le rôle de support une fois la configuration de départ installée. Le regard accompagne le déploiement de la forme du signe jusqu'à la fin.

2.2.1.3 Instanciation de signe *« comme ça »*

[Cuxac 2000] souligne que le regard permet de créer des instances d'un signe en même temps que sa réalisation. Ceci signifie que le signe créé est à caractère standard [UN ARBRE]. Alors que le *« comme ça »* représente un signe spécifié [CET ARBRE LÀ]. [Cuxac 2000] qualifie l'instanciation de signe d'*« illustrative »* contrairement à la réalisation de signe standard qu'il qualifie de *« non illustrative »*. De plus, [Cuxac 2000] souligne que le regard est le garant du passage du *« non illustratif »* vers l'*« illustratif »*. On parle ainsi d'une condition qui permet ce passage, celle de fixer tout ou une partie du signe.

1. Les corpus que nous avons traités ont été acquis à la fréquence de 25 images par seconde. La durée d'une image correspond donc à 0.040 seconde

2.2.2 Transfert personnel

Ce modèle décrit les étapes de prise de rôle. Quand le signeur prend le rôle d'une personne de l'énoncé [Cuxac 2003], son regard quitte l'interlocuteur. Ensuite, un clignement des yeux marque le début de transfert. Ainsi, le signeur ne fixe pas l'interlocuteur pendant toute la phase de prise de rôle puis vient poser son regard de nouveau sur son interlocuteur, après un clignement des yeux, pour marquer la fin du transfert.

2.2.3 Changement de références

C.Cuxac souligne que le regard assure le changement de références dans un discours en Langue des signes. De sa part, [Leroy 2010] s'appuie sur des exemples de dialogue enseignant – élève en LS pour montrer que le regard assure le maintien de l'interaction entre locuteur et interlocuteur et l'organisation d'échanges entre plusieurs signeurs. [Lejeune 2004, p. 60-64] souligne que le regard assure le passage d'une séquence narrative à une reprise de dialogue avec l'interlocuteur. Par ces exemples, nous déduisons que le regard joue un rôle linguistique qui dépend uniquement des actants [2].

2.2.4 Le déictic ou le référencement

Le déictic est défini comme étant la fonction qui permet de mettre le « *focus* » sur un locus par le regard ou la main [Risler 2005]. Cette fonction est présente dans les procédés :

D'activation d'un locus :Il s'agit de l'activation d'une zone de l'espace localisée antérieurement dans le discours. [Lejeune 2004, p. 60-64], [Risler 2005], [Rinfret 2009] et [Meurant 2003] confirment que le regard active et marque des zones de l'espace de signation localisées antérieurement dans le discours.

De construction d'un locus : Il s'agit de l'activation d'une zone de l'espace dans le but de lui associer un signe.

De désignation des actants de l'énoncé :la désignation des pronoms [JE], [TU], etc.

De signes en mouvement : dans le déploiement de gestes par la tête et / ou le buste pour marquer les zones repères (départ et destination du signe en mouvement).

Pour illustrer la présence de référencement dans le procédé de signe en mouvement, nous avons choisi l'exemple [DONNER] dont les gestes déployés pour le réaliser désignent l'actant (celui qui réalise l'action), le destinataire ou le patient (celui qui reçoit l'action) et l'objet de l'action résultante (voir l'exemple 2.1 dont l'objet donné est un verre).

2. Les acteurs du discours peuvent être des personnes, des objets, une situation, etc.

FIGURE *2.1 – Signe en mouvement [DONNER] signifiant dans cet exemple « donner un verre ». La configuration de la main reprend la forme d'un verre. Les actants du signe [DONNER] localisés par les positions de départ et de destination du mouvement de la main.*

2.2.5 Conclusion

Nous avons présenté les procédés linguistiques de construction et d'instanciation de référence qui décrivent les gestes en lien avec l'espace de signation. Nous avons souligné que la fonction qui permet de faire référence à un élément du discours, que ce soit un actant de l'énoncé ou une entité localisée dans l'espace de signation, représente une partie de ces procédés. Dans la construction de référence, le signeur fixe une zone de l'espace ou un locatif (la main dans un transfert situationnel) avant la localisation d'une entité du discours *« locus »*. Nous avons noté que cette fonction est également présente dans : 1) la réalisation de signes en mouvement quand le signeur fait référence aux zones représentant l'actant et le patient du signe en mouvement ou de l'action, dans 2) les pronoms quand le signeur désigne l'interlocuteur (en signant [TU] par exemple).

2.3 Les catégories du référencement

Nous choisissons d'étudier le *« Référencement »* qui représente la désignation d'un locus de l'espace de signation associé à un signe réalisé à un instant antérieur. Nous nous proposons d'étudier ses formes gestuelles qui permettent l'activation d'une zone ou la réutilisation d'un locus de l'espace. Dans un premier temps, nous allons énumérer les catégories linguistiques de *« Référencement »*.

2.3.1 Les catégories linguistiques

Nous présentons les variantes de référencement étudiées en quatre catégories telle qu'elles ont été classées par [Bras 2004] :

- Le pointage ou le pointé [3] (figure 2.2). .
- Le signe localisé.
- Le signe en mouvement.
- Le signe locatif.

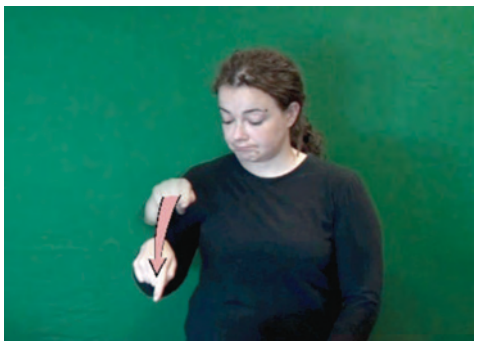


FIGURE *2.2 – Un pointage avec un index tendu*

2.3.1.1 Le pointé

Les exemples de la figure (2.3) illustrent deux configurations manuelles du pointé ; index tendu  et main plate  [B]. Les deux gestes tels qu'ils sont illustrés dans (2.3) sont marqués par un arrêt pour marquer la zone désignée. [Bras 2004] souligne que le pointé peut désigner un simple point ou une portion de l'espace. Peut-on dire qu'il y a une relation entre la forme conceptuelle [4] d'un locus désigné et la configuration du pointé ?

3. Le pointage est aussi appelé dans certains ouvrages *« le signe pointé »* ou *« le pointé »*
4. telle que le signeur l'imagine

FIGURE 2.3 – *Un pointé avec : a) main plate, b) index tendu*

Plusieurs linguistes se sont penchés sur la relation entre configuration manuelle de pointé et forme conceptuelle d'un locus dont [Engberg-Pedersen 1986] pour la DSL, [Padden 1983] pour l'ASL et [Deuchar 1984] pour la BSL. Ils partagent le fait que si l'endroit pointé est un point de l'espace, la configuration de la main est :  et si l'endroit pointé une zone plus étendue qu'un point, la configuration de la main est :  .

D'autre part, [Rinfret 2009] introduit le paramètre d'accessibilité qui représente la manière dont un locus est pointé. Concrètement, le terme d'accessibilité signifie l'amplitude et le nombre de gestes déployés lors d'un pointé. [Rinfret 2009] quantifie le paramètre d'accessibilité de *« fort »* :

1. si le geste déployé est caractérisé par un déplacement important ;
2. si le nombre de gestes déployés est important.

[Rinfret 2009] quantifie de *« faible »* dans les cas opposés à (1) et/ou (2). L'auteur a pu mettre en relation la *« faible »* ou *« forte »* accessibilité avec la forme conceptuelle d'un locus. Une faible accessibilité est liée au fait qu'un locus est un point. Plus encore, [Rinfret 2009] postule une relation entre l'accessibilité d'un locus et le temps de son énonciation [5] On retient que le référencement permet de déterminer le degré d'accessibilité à un locus sur les plan spatial et temporel.

Point de vue fréquence du pointé, [Fusellier-Souza 2004] a recensé le nombre d'occurrences de pointés dans un corpus en langue des signes émergente selon leurs fonctions linguistiques et en a conclu que la sous-variante linguistique du pointé manuel – *« La référence spatiale »* est la plus productive parmi les six types de fonctions du pointé identifiées dont celles qui ont également une relations avec l'espace de signation : 1) Le signe localisé, 2) la référence locative – *« comme ça »*.

5. temps d'énonciation : localisation du locus dans le tempss par rapport au début de l'énoncé.

2.3.1.2 Le signe localisé

Il s'agit de la reprise d déploiement de signes localisés préalablement dans l'espace suivi d'un pointé. [Fusellier-Souza 2004] énonce trois interprétations de signes localisés :

- Valider les propos de l'interlocuteur ;
- Marquer la fin de l'énoncé ;
- Le pointé réalisé par la main dominante vers la main dominée réalisant ce signe [6].

Cette dernière catégorie apparaît comme une classe à part entière dans la catégorisation de [Bras 2004].

2.3.1.3 Le signe en mouvement

En langue des signes française de Belgique (LSFB), [Meurant 2003, p.9-10] évoque le déploiement de gestes réalisés par le buste. Ces gestes suivent l'orientation du mouvement manuel dans un signe en mouvement. Le mouvement localisé qui représente une sous-partie de la réalisation d'un signe en mouvement permet de marquer ses zones de départ et d'arrivée. Les zones de départ et de destination (actant et patient) peuvent correspondre à deux locus ou un locus et une personne [Meurant 2007, p.408-409].

Exemple dans la phrase signée *« Il lui donne la pomme »* – [POMME] [IL] [DONNE] [LUI] – le mouvement du buste s'oriente vers un locus de départ [IL] puis un locus d'arrivée [LUI]. Outre les mouvements du buste, [Parisot 2003] note qu'en langue des signes québécoise (LSQ), le regard participe également dans la mise en relation entre la zone de départ et celle d'arrivée du signe en mouvement et en particulier pour activer ces zones de l'espace.

2.3.1.4 Le signe locatif

Nous reprenons la fonction de transfert de situation décrit la localisation d'un signe $S1$ dans l'espace par rapport à la position d'un autre signe $S2$ associé préalablement à une zone de l'espace. Le signe locatif se réalise par 1) la main dominée qui en fin de réalisation de $S1$ localise la zone pendant un certain temps et 2) par le regard quand celui-ci fixe le locatif stable (main dominée immobile).

2.3.1.5 Constatations

Outre le fait que la notion de référencement est présente dans le pointé, le signe localisé et le signe en mouvement, nous remarquons que le référencement montre le degré d'accessibilité d'un locus pointé selon les deux facteurs : l'espace et le temps. Ceci met en avant une relation spatio-temporelle entre la localisation spatiale d'un locus et son référencement. Nous pensons que ce constat est intéressant pour automatiser l'analyse de la notion de référencement car les critères de représentations (le temps et l'espace) sont mesurables et permettent d'exploiter au mieux l'information gestuelle. De même pour le constat établi à propos de la relation entre configuration manuelle du signe [POINTÉ] et la forme conceptuelle du locus.

Nous en retenons que les paramètres à prendre en compte pour la description informatique du référencement sont sur deux niveaux : spatio-temporel exprimant l'organisation

6. Dans ce cas, le terme *« locatif »* est associé à la main

de l'espace de signation, et gestuel exprimant les actions déployés dans le but de modifier le contenu de l'espace ou de l'enrichir par des informations supplémentaires.

2.3.2 Les variantes gestuelles

[Cuxac 2000, p.282] affirme que le regard et le pointé par la main participent conjointement à l'organisation du contenu de l'énoncé dont le référencement. [Fusellier-Souza 2004] ajoute que l'accord direction du regard - pointé dans le référencement se réalise au moment de la construction de référence. Nous avons mentionné dans la section précédente l'accord regard – pointé omniprésent dans l'interaction enseignant - élève [Leroy 2010]. Les exemples (2.3 a et b) illustrent un accord regard - tête - main et regard - tête - buste - main. Dans les deux exemples, les composants corporels sont orientés vers un même locus. La multi-linéarité du référencement a été souligné par [MacLaughlin 1997] qui énonce que, en langue des signes américaine, l'inclinaison de la tête est aussi impliquée dans le marquage d'un locus, alors que [Winston 1995] met l'accent sur la rotation de la tête. De sa part, [Engberg-Pedersen 2003] parle de l'orientation de tout le corps. En langue des signes québécoise, [Parisot 2006] confirme que l'inclinaison latérale du buste (ou le *« tronc »*) est autant concernée par l'activation d'un locus que le regard et la main (le pointé). [Rinfret 2009] établit une analyse fine des types de mouvement du tronc et affirme que l'inclinaison latérale du tronc marque l'agent réalisant une action alors que le regard permet de marquer le patient (celui qui subit l'action).

D'autre part, [Fusellier-Souza 2004] attribue au procédé répétitif de pointé la fonction d'énumération dans le cas où la main dominante pointe la main dominée.

2.4 L'espace de signation

Le signeur utilise l'espace pour réaliser des signes par les mains, des mouvements corporels comme la rotation de la tête et le balancement du buste. L'usage de l'espace est représenté par plusieurs fonctions linguistiques liées à la nature de l'espace (linguistique ou physique). [Cuxac 2005] définit l'espace de signation comme étant l'espace qui permet la réalisation de messages dont ceux de structures fonctionnelles tels que les fonctions de transferts abordées dans (2.4.5) (détaillées dans la section suivante.)

2.4.1 Décomposition de l'espace de signation

[Risler 2005] présente une décomposition de l'espace de signation en trois niveaux. L'espace de signes lexicaux regroupe les signes construits par les mouvements manuels. L'espace syntaxique regroupe les relations de nature spatiale comme la forme et les mouvements des signes. L'espace topologique illustre les relations spatiales entre les entités linguistiques localisées *« loci »* ou entre signeurs et *« loci »*.

2.4.2 Construction de l'espace de signation

[Risler 2005] et [Rinfret 2009] soulignent que l'association espace-signe se fait selon plusieurs modalités manuelles et/ou non manuelles. [Risler 2005] mentionne que plusieurs articulateurs participent à la construction des espaces lexical, syntaxique et topologique. L'espace de signes lexicaux est construit par les mouvements manuels. L'espace syntaxique est établi par les mouvements qui ont permis la construction de signes alors que l'espace topologique est construit par le regard et l'index pointés vers les signes localisés dans l'espace.

Terminologies :

Le *« geste co-verbal »* est usuellement défini comme étant un geste qui accompagne la parole. Dans le cas des Langues des Signes, ce sont les mêmes articulateurs qui sont utilisés pour produire des gestes et des mouvements, quelque-soit leur nature, qu'ils soient porteurs de sens ou pas. Ainsi, il est difficile de trouver des critères permettant de définir ce que seraient des gestes "verbaux" ou "non-verbaux" et la notion de "verbal" semble difficilement transposable dans le cas des LS. [Cuxac 2005] souligne ce caractère flou de la frontière entre la syntaxe verbale et la syntaxe non-verbale. De sa part, [Risler 2005] met l'accent sur l'importance de l'espace topologique dans la distinction de la syntaxe verbale de la syntaxe non-verbale.

2.4.3 Synthèse

L'espace de signation est le résultat de fusion d'interprétations d'unités gestuelles. La caractérisation de l'espace revient à l'étude de gestes impliqués dans sa construction dont les signes de pointé, les signes localisés et les signes en mouvement.

Nous nous proposons de mener une étude sur les réalisations gestuelles de ces variantes de signes qui font partie du référencement comme nous l'avons mentionné dans (2.3.1).

2.4.4 Modèles linguistiques gestuels

Certains rôles exprimés par le regard font intervenir, en plus des cibles pointées, des clignements des yeux, des haussements de sourcils, des rotations de la tête plus ou moins marquées, des gestes manuels et des balancements du buste.

Dans le cadre de l'étude de la LSF dans l'interaction enseignant – élève, [**?**] a mis en évidence la présence du couple regard-pointage omniprésent dans l'interaction. [**?**] a également constaté que la gestion des tours de parole et l'incitation à la prise de parole se manifestent par :

- le regard renforcé par un mouvement vers l'arrière de la tête et/ou du menton ;
- le regard fixe et un menton relevé ;
- une mimique interrogative et un léger mouvement du menton vers le bas ;

Dans son étude sur les gestes de désignation de pronoms, [Meurant 2005] a souligné la pertinence de la coïncidence pointage-regard dans la différenciation entre pronoms.

Par exemple, les pronoms [TU] et [IL] se différencient uniquement par la coïncidence pointage-regard.

2.4.5 Le regard et l'espace

Le terme espace figure dans les fonctions linguistiques exprimés par le regard que nous avons évoqué dans la section (2.3.1). L'espace représente une cible du regard dans certains de ces fonctions. Nous qualifierions la cible par *cible directe* où l'espace est impliqué dans la construction de certains des fonctions linguistiques d'une manière directe. Dans les autres cas, l'espace est évoqué d'une manière indirecte. Par exemple, dans la structure de *« dire tout en montrant »*, le regard porte sur la main qui localise un signe dans l'espace. Dans ces cas, nous qualifierons l'espace de *cible indirecte* où l'espace est impliqué dans le rôle d'une manière indirecte. L'implication indirecte de l'espace est présente également dans le maintien de la communication, en désignant par le regard un objet ou une personne physique appartenant à la scène réelle.

L'objectif étant d'identifier des catégories des fonctions linguistiques liées au regard, nous avons choisi le classement par *« type »* de relation : Regard – espace. Nous avons choisi d'attribuer deux valeurs pour chaque *« type »* de relation : directe et indirecte telle que nous l'avons mentionné. Le diagramme de la figure (2.4) propose une formalisation de cette méthode de catégorisation. Nous avons identifié dans la catégorie *« relation directe »* des sous-catégories de fonctions linguistiques distinguées par le nombre de cibles comme cela a été illustré dans les figure (2.5 et 2.6). Ce diagramme permet de situer les fonctions linguistiques faisant partie du référencement par rapport aux fonctions linguistiques, en particulier, celles où le regard est impliqué.

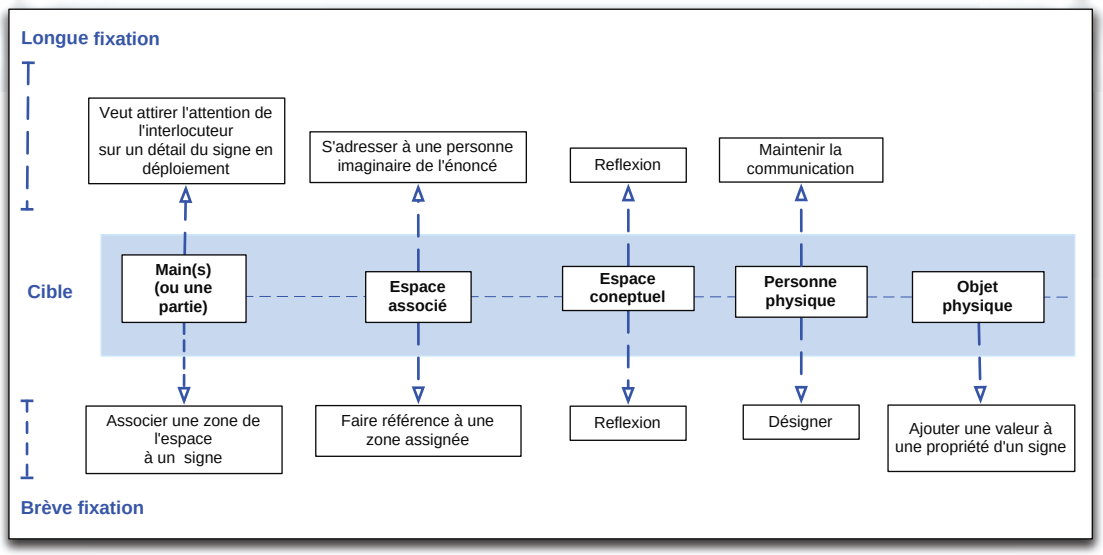


FIGURE 2.5 – *Intentions du regard fixant une unique cible*

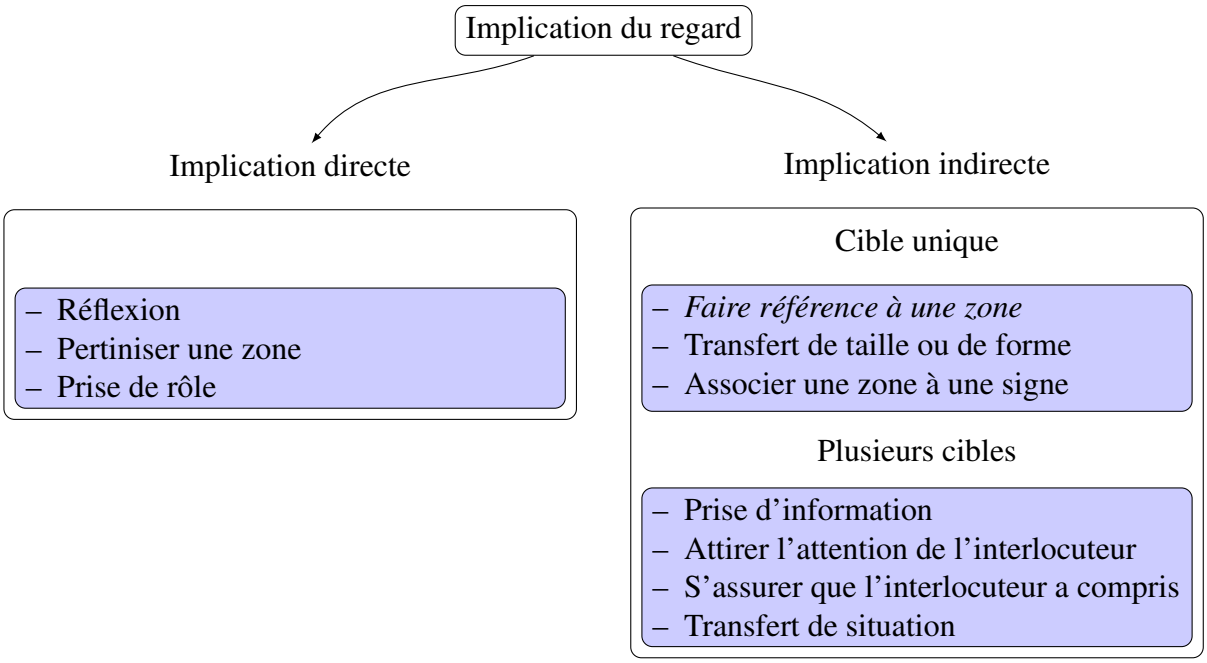


FIGURE *2.4 – Rôles du regard par : 1) Implication directe du regard et 2) Implication indirecte du regard. Représentation de l'implication indirecte sous forme de deux sous classes : Cible unique et Plusieurs cibles*

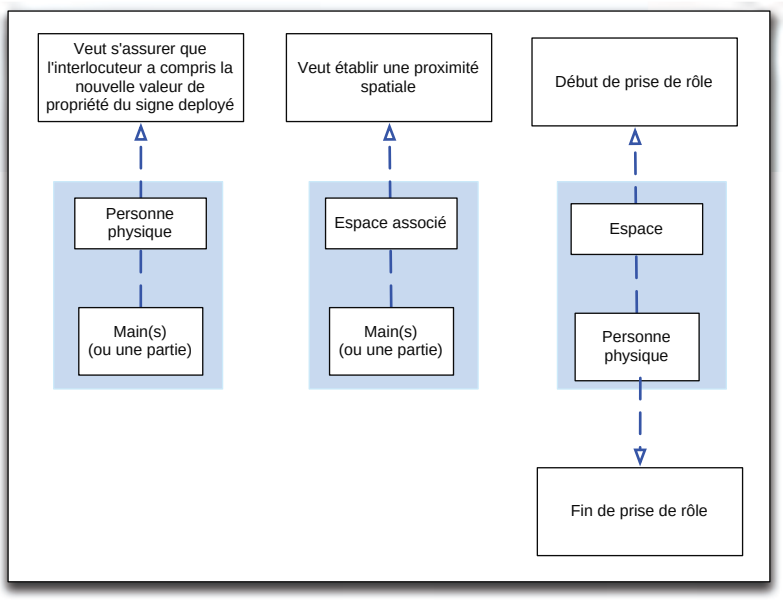


FIGURE *2.6 – Intentions du regard fixant deux cibles consécutives*

2.5 Retour sur objectifs

Ainsi, le référencement est un concept linguistique complexe en terme de représentation car plusieurs fonctions linguistiques y sont liées (de manières différentes). A chacune de ces fonctions linguistiques correspondent plusieurs combinaisons gestuelles.

D'autre part, nous avons constaté des difficultés dans la perception de l'orientation du regard. En effet, le format d'images utilisé – basse résolution – dans l'acquisition de corpus vidéos rend fastidieuse la tâche de déterminer dans quelle direction regarde le signeur que ce soit manuellement (annotation) ou automatiquement (détection automatique de l'angle d'orientation du regard). Ces constatations nous ont amenés à choisir une variante linguistique de référencement :

- exprimée par plusieurs combinaisons gestuelles y compris le regard afin de mettre en place un modèle robuste qui tient compte de la perception du regard en se servant des informations véhiculées par les gestes de la tête, des mains et du buste,
- la plus productive afin d'avoir une base représentative de données réelles.

En se basant sur le fait que les énoncés en langue des signes comportent des signes, dans la plus part de cas, destinés à y faire référence et sur le fait qu'un référencement inclut plusieurs fonctions linguistiques et est réalisé de plusieurs manières (manuelles et / non manuelles), nous avons choisi d'étudier le concept de référencement. Pour cela, nous étudierons les variantes linguistiques qui incluent la notion de référencement et qui

ont été énumérées dans (2.3.1). Dans l'objectif de représenter le référencement, nous étudierons les réalisations gestuelles. Nous proposerons deux types de représentation informatique :

- Les combinaisons gestuelles de référencement en se basant sur l'ordre dans lequel les gestes sont réalisés,
- La disposition spatiale d'un locus par rapport à un autre.

Nous avons plusieurs choix de paramètres à mesurer pour caractériser l'aspect temporel et spatial du geste de référencement et des entités référencées.

L'aspect temporel regroupe les décalages temporels entre mouvements, la durée et la fréquence d'un geste, etc. L'aspect spatial regroupe la position de l'entité référencée (absolue ou relative), la direction, la position et le déplacement de la composante corporelle en phase de référencement, etc. Nous avons choisi d'étudier les positions relatives de l'entité référencée ainsi que la direction de la tête en phase de référencement. Dans le chapitre (4), nous présenterons les paramètres choisis en fonction des données des corpus dont nous disposons.

Le but final étant de mettre en place une méthode de reconnaissance de structures de référencement, nous considérons que le cahier de charge de la méthode se basera sur les invariants cités ci-dessus. De ce fait, la reconnaissance de structures de référencement sera reformulée en une méthode de segmentation d'enchaînements de gestes pointant ou pas vers le même endroit et combinés d'une manière spécifique. Par quel(s) paramètre(s) peut-on caractériser la convergence des gestes (vers un même locus) réalisés au cours d'un référencement ?

Dans la suite du mémoire, nous chercherons à

- affiner la notion de pointé en manuels et non manuels,
- définir un modèle de combinaison gestuelle en terme d'ordre chronologique de réalisation de gestes,
- caractériser la relation geste - locus.

Nous étudierons la production gestuelle de référencement et présenterons les modèles que nous avons construits. Ces modèles nous permettront de nous affranchir des paramètres variables dont l'emplacement de début et de fin des composantes corporelles.

Pour cela nous allons décrire les corpus que nous avons utilisé pour la modélisation du pointé.

CHAPITRE 3

Corpus

Outre l'importance du type et du volume des données contenues dans le corpus, les données devraient être diversifiées et donc représentatives de la complexité de la langue des signes, en termes de variantes gestuelles, tout en restant dans un champ lexical limité pour simplifier son exploitation. Dans ce chapitre, nous décrirons les corpus traités afin de mettre en avant les caractéristiques temporelles et spatiales des gestes de référencement.

Sommaire

3.1 Cahier des charges

3.1.1 Objectifs

Le corpus devrait permettre de :

- Extraire les caractéristiques gestuelles du référencement à partir de modèles linguistiques.
- Traduire les caractéristiques en représentations géométriques $3D$
- Déduire des modèles d'aspect ($2D$) exploitables en traitement de vidéos de langue des signes.
- Evaluer les modèles d'aspect en les appliquant sur des vidéos uniquement.

Les modèles gestuels construits devraient répondre aux questions suivantes :

Quelles sont les propriétés d'un geste de référencement ?

Quels sont les paramètres qui permettent de les exprimer ?

L'objectif d'un modèle réalisable est d'adapter les propriétés du référencement à un programme de détection automatique de référencement dans une vidéo.

3.1.2 Comment ?

Afin de construire des modèles représentatifs de la réalisation de référencement, nous aurons besoin d'un corpus composé de vidéos annotées et de données tri-dimensionnelles des composantes corporelles en mouvement. L'annotation devrait permettre d'assister le programme de détection de structure de référencement en lui fournissant les informations suivantes en amont :

- le signe localisé dans l'espace de signation.
- le référencement d'une zone de l'espace de signation.
- le type de référencement : manuel, non-manuel.
- le type de zone référencée : un locus, un groupe de locus.

La transcription des vidéos devrait comporter les interprétations linguistiques suivantes :

- le signifié du signe réalisé (gloses).
- la localisation de signes dans l'espace de signation.
- les référencements manuels et non-manuels.
- les cibles de référencement : locus seul, groupé, etc.

Afin d'obtenir des modèles d'aspect, nous aurons besoin d'une méthode de passage $3D \Rightarrow 2D$ qui permettent de conserver les informations pertinentes au repérage de structures de référencement. L'évaluation des modèles d'aspect requiert des vidéos enregistrées dans des conditions particulières qui permettent un repérage non biaisé des zones d'intérêt dans une image telles que la tête et les mains.

3.1.3 En résumé

Nous avons établi le cahier de charge des corpus. Trois types de corpus composés de vidéos, d'annotations textuelles, de capture de mouvement et du regard. Le scénario du corpus de modélisation sera marquée par une utilisation fréquente de l'espace de signation. Les vidéos de corpus d'exploitation et d'évaluation seront enregistrées dans des conditions particulières qui faciliteront le repérage des mains et de la tête. Les conditions d'enregistrement incluent le port d'un habit à manches longues et un fond uniforme derrière le signeur. Les données de capture du regard seront utilisées dans le corpus d'exploitation afin de compenser l'absence de données de capture de mouvement.

Pour pouvoir mettre en oeuvre notre système de détection, nous aurons besoin d'un corpus qui répond aux besoins exprimés dans le tableau (3.1) ce qui n'est pas le cas en réalité. Nous ne pouvons pas disposer de corpus fournissant à la fois des données $3D$ des composantes corporelles, des données du regard et des vidéos enregistrées dans des conditions particulières car la présence de marqueur de capture de mouvement rend difficile le traitement automatique des vidéos. Nous devrons donc utiliser des corpus différents pour chaque étape : Modélisation, exploitation et évaluation.

1. Capture de mouvement 3D
2. C :Avec contraintes d'enregistrement

Objectifs	Vidéo	Mocap [1]	Regard	Transcription	Conditions
Modélisation	✓	✓	✓	Textuelle	Utilisation marquée de l'espace de signation
Exploitation	✓- C [2]		✓	Textuelle	
Evaluation	✓- C				Enonciation naturelle et continue Champs lexical restreint

TABLE *3.1 – Cahier de charge du corpus à mettre en place*

3.2 Corpus de modélisation

Nous avons analysé deux corpus composés de données de capture de mouvement et un troisième corpus composé uniquement de vidéos annotées. Notre choix de système de capture de mouvement s'est basé sur le fait que la capture de mouvement permet d'enregistrer des corpus qui rendent compte de la production gestuelle avec un minimum de biais, ceci est également la constatation de [Elliott 2007]. Le corpus composé uniquement de vidéos annotées a été utilisé pour étudier les propriétés temporelles des fonctions linguistiques liées au référencement.

3.2.1 Etude temporelle : corpus *« Websourd »* [3]

Nous avons opté pour un corpus de brèves d'actualités en langue des signes française (LSF) réalisés à Websourd dans l'objectif de diffuser les actualités de différents domaines (ex. La santé, la politique, l'économie, etc.). Nous avons choisi ce corpus car les équipements d'enregistrement sont non intrusifs ; le signeur n'a pas de contraintes ni sur la manière de signer ni sur les habits qu'il porte. De plus, les sessions d'enregistrement sont courtes et concises.

Afin d'étudier de plus près les propriétés spatiales des gestes de référencement, nous avons opté pour deux corpus dont les données sont variées (3D et vidéos). Sur le plan des scénarios, les sessions d'enregistrement contiennent plusieurs variantes du référencement et présente une utilisation fréquente de l'espace de signation.

3.2.2 Etude spatiale

Les corpus de modélisation spatiale ont été réalisés avec une caméra de scène qui filme une vue de face du signeur, son cadrage large permet une capture globale sans

3. WebSourd est une entreprise qui a pour vocation la mise en œuvre de services d'accessibilité pour les personnes sourdes et le développement d'outils et métiers utilisant internet et les NTIC et qui favorisent leur indépendance et leur citoyenneté.

imposer trop de contraintes sur les mouvements du signeur. Le système *« VICON »* de capture de mouvement 3*D* est composé de caméras infra-rouges et des marqueurs réfléchissants disposés à la surface des articulations et sur le visage du signeur. La taille faible des marqueurs, de l'ordre de 3 à 10 mm de diamètre, permet de suivre la posture, la déformation du visage et les configurations manuelles du signeur. Les différents angles de vue des caméras infra-rouges utilisées pour la capture de mouvement permettent de compenser l'absence de données de certains marqueurs. Cette technique permet d'estimer la position des marqueurs avec une erreur moyenne de 0,5 millimètre[Viel 2003].

Cependant la présence de marqueurs sur le visage et sur les doigts introduit une gêne pour la production de signes comportant un contact main-visage ou main-main, en particulier dans le cas de *« contact-glissé »*.

La mise en place du corpus 3*D* est coûteuse compte tenu du temps de la pose des marqueurs sur le corps, le visage et les mains du signeur (Voir figure 3.1). Le signeur est piloté par un intervenant sourd hors champ d'acquisition qui joue le rôle du second interlocuteur dans les phases de dialogue et qui peut jouer un rôle de souffleur dans les phases de narration. Pour les sessions de tournage, le signeur s'appuie sur des supports visuels projetés (au lieu de texte écrits) afin de limiter l'influence de la langue française. Cette méthode permet au signeur d'énoncer directement les concepts dans sa langue.

Pour chacun des deux corpus, un protocole d'acquisition a été mis en place. Il contient entre autres une méthode de synchronisation temporelle des données enregistrées. Le signeur commence et termine sa production par un *« clap »* manuel qui permettra de synchroniser ultérieurement la vidéo et les données de capture de mouvement.

3.2.2.1 Corpus *« SignCom »* [4]

Un corpus en langue des signes française (LSF), a été réalisé dans le cadre du projet ANR *« SignCom »*. Son thème principal porte sur les recettes, plus précisément des recettes de galette, de cocktail et de salades avec un champ lexical différent pour chaque sous-thème. L'espace de signation représente le plan de travail, les objets et les ingrédients utilisés. Dans chaque scénario présentant une recette, on peut modifier les dispositions spatiales des ingrédients ou l'ordre dans lequel on les fait intervenir. On peut ainsi limiter le champ lexical à une liste d'objets et d'ingrédients et des action (préparations de plats), tout en exploitant les variantes possibles des signes au niveau de la forme et de la taille (ex. grand / petit).

3.2.2.2 Corpus *« Marqspat »* [5]

Le corpus Marqspat – en langue des signes québécoise (LSQ), a été enregistré dans le cadre du partenariat franco-québécois *« Marqspat »* dans lequel se développent différents projets qui portent sur le thème de marquage spatial dans les langues des signes française,

4. Le projet Sign'Com vise à améliorer la communication entre agent réel et un agent virtuel. http ://www-valoria.univ-ubs.fr/signcom/fr/. Le projet se base principalement sur des spécifications pour la synthèse.

5. http ://www.irit.fr/marqspat/index.html

américaine et québécoise. Il s'agit d'un corpus d'élicitations guidées et libres où le signeur répond à des questions posées par l'interlocuteur en face de lui sur le contenu de vidéos projetées et sur des expériences personnelles similaires. Les trois vidéos projetées sont des séquences de scène où deux personnes sont dans :

- une salle d'attente,
- un atelier de peinture,
- un entretien d'embauche.

La figure (3.1) représente une prise de vue de l'interlocuteur et des scènes projetées lors d'une session d'enregistrement.

FIGURE *3.1 – Prise de vue de l'interlocuteur et des scènes projetées*

Le système d'acquisition inclut un équipement de capture du regard *« FaceLab »*. Nous avons établi un comparatif d'outils de capture du regard afin de justifier notre choix pour la méthode de capture implémentée dans cet équipement (Voir Annexe A). Il s'agit d'un équipement non invasif composé de deux caméras et d'un émetteur infrarouge. Il fournit sous forme de données tri-dimensionnelles, l'orientation du regard et celle de la tête. Il fournit également un enregistrement du cadre général de la scène. Les images contenues dans les enregistrements vidéos sont numérotées au fur et à mesure que nous défilons l'enregistrement.

A l'aide de l'enregistrement du cadre général de la scène fourni par FaceLab et les vidéos fournies par la caméra de scène, on peut déduire le décalage temporel qui existe

entre la vidéo du cadre général de la scène fournie par Facelab et celle fournie par la caméra. Cette méthode a été utilisée manuellement pour chaque session d'enregistrement.

Pour chacun des corpus, nous avons considéré deux séries de vidéos, la première pour la construction de modèles **(S1)** et la seconde pour les tester **(S2)**.

3.2.3 Annotations

Nous avons eu besoin d'annoter les vidéos des corpus de modélisation dans le but, d'une part, d'enrichir les modèles linguistiques de référencement par des mesures qualitatives et quantitatives et, d'autre part, dans le but d'évaluer les fonctions linguistiques de référencement détectées automatiquement. L'annotation est une tâche, certes fastidieuse, mais indispensable pour l'élaboration de modèles linguistiques de référencement.

Le logiciel Elan [6] a été utilisé pour annoter manuellement les éléments pertinents à l'analyse de référencement dans les vidéos (LSF) et (LSQ) ainsi que pour visualiser les résultats de segmentation automatiques de séquences de référencement dans l'objectif de tester les modèles construits. Nous avons fait appel à des experts en langue des signes française et québécoise pour :

- Délimiter les événement de référencement dans la première série de vidéo (S1).
- Evaluer les annotations automatiques dans la seconde série (S2).

3.2.3.1 Annotation du corpus *« Marqspat »*

L'annotation du corpus a été effectuée par l'équipe de recherche québécoise en langue des signes [7]. Le corpus a été annoté sur plusieurs niveaux de fonctions linguistiques telle que la fonction de localisation. L'annotation de l'espace de signation consiste à décrire sa composition au fur et à mesure qu'un locus est créé. Nous avons utilisé l'annotation des question posées par l'intervenant et rapportée dans la piste *« Question »* pour comprendre le thème général de l'énoncé. Nous avons repéré les réponses du locuteur sous forme de gloses annotées dans la piste *« MD »*. La figure (3.2) illustre un exemple d'annotation d'une question et de la réponse correspondante sous forme de gloses.

La piste *« MD »* concerne les signes manuels réalisés par la main droite y compris le signe de pointé comme on le constate dans l'exemple de la figure (3.2). Les deux pistes nous permettent de repérer le signifiant du signe pointé et les pointés correspondants. Pour pouvoir localiser les zones associées aux cibles pointées, nous avons une troisième piste *« Localisation MD »* qui représente les segments de localisation de signes effectuée par la main droite. Chaque zone spatiale est étiquetée par une lettre en minuscule s'il s'agit d'un locus sinon par une lettre en majuscule s'il s'agit d'un ensemble de locus. Si dans un même segment, un locus est référencée plusieurs fois, il aura plusieurs labels numérotés (*x1*,*x2* etc.). Les indices placés devant les labels (ex. *x**1***) représentent les signes associés à ce locus. L'exemple de la figure (3.2) illustre l'association du signe

6. Elan (EUDICO Linguistic Annotator) est un outil de création, d'édition, de visualisation et d'annotation de données vidéos et paroles. Il a été développé à l'institut Max Planck à Nimègue au Pays-Bas.

7. *« Groupe de recherche sur la langue des signes québécoise et le bilinguisme sourd »* qui pilote le projet *« Marqspat »*

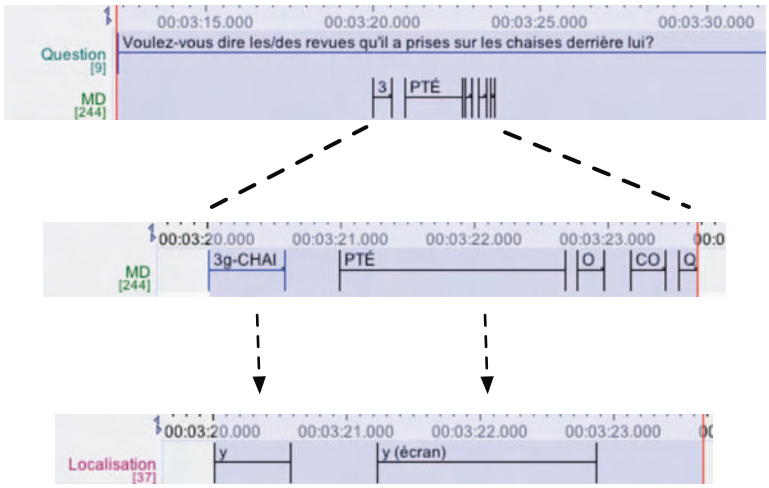


FIGURE *3.2 – Exemple d'annotation d'un segment de question / réponse*

[CHAISE] au locus étiqueté *« y »* puis d'un signe de pointé vers ce même locus. Les trois pistes *« Regard »*, *« Tête »* et *« Tronc »* délimitent les segments de référencement non-manuels. Les valeurs correspondantes sont attribuées selon le même protocole que celui de la piste *« Localisation »*.

Constatations : Nous avons mesuré le taux d'occurrence des référencements réalisés par la main droite, le regard et la tête dans une partie du corpus. Les référencements apparaissent en une très faible proportion : sur une durée totale de 16 minutes et 4 secondes, nous avons 15.3% de cette durée annotée sous forme de valeurs de mouvements manuels dont 16.4% de référencement. D'autre part, nous avons 35,8% de la durée totale qui a été annotée sous forme de valeurs de direction du regard dont l'interprétation de 13.4% des données données transcrites sont des référencements.

3.2.3.2 Annotation du corpus *« Websourd »*

L'annotation du corpus *« Websourd »* a été réalisée par notre équipe avec l'aide d'une experte en (LSF). Le schéma de l'annotation contient des référencement manuels et non-manuels. Nous avons choisi de remplacer l'annotation de la rotation du buste qui semblait pertinente au référencement par celle des épaules car leurs mouvements sont mieux perceptibles. Nous adoptons l'appellation *« tronc »* pour désigner à la fois les mouvements des épaules et ceux du buste [Parisot 2011].

Annotation des signes manuels : L'annotateur prend connaissance du contenu de l'énoncé, puis délimite les signes manuels sous forme de gloses. La stratégie de délimitation consiste à repérer la stabilisation de la configuration manuelle et marque le début du signe. L'annotateur indique la fin du signe en repérant des indices de début de signe qui concernent le mouvement ou la configuration de la main réalisant le signe. Il

repère une décélération du mouvement ou un changement de configuration de la main réalisant le signe. Ainsi la fin du signe correspond à l'image qui la précède. L'exemple de la figure (3.3) est un exemple de l'annotation du signe [LIEU]. Les figures 3.3.(1) et 3.3.(5) représentent des exemples de pré et post stabilisations de la configuration manuelle. Par ailleurs, nous avons noté qu'il existe d'autres stratégies de délimitation de signes manuels. Par exemple, les critères de [**?**] sont explicités sous forme d'un algorithme ayant comme paramètres : 1) la position de fin du signe précédent (si commun avec le signe actuel), 2) la position de départ du signe actuel (si elle est fixe) et 3) la direction du mouvement du signe.

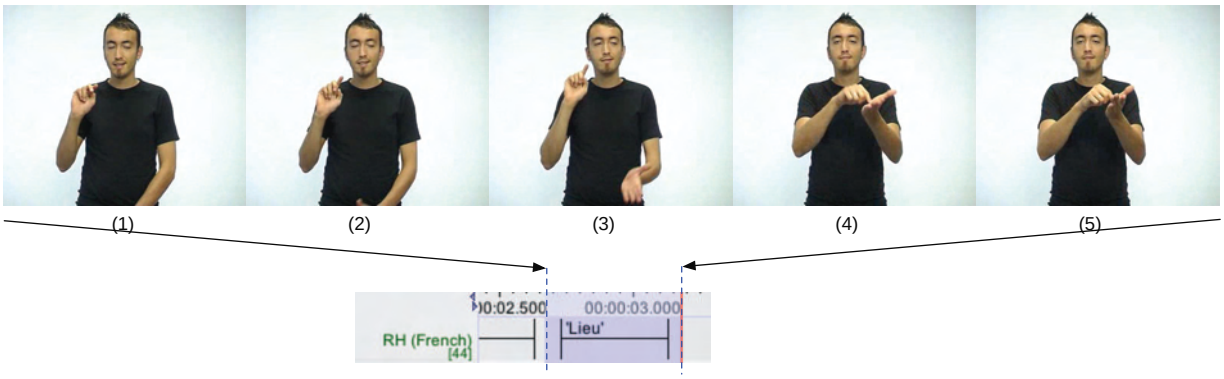


FIGURE *3.3 – Annotations de signes manuels*

Annotation des mouvements : Il s'agit de l'annotation qualitative des rotations et des déplacements :

- de la tête : selon les axes vertical, horizontal et celui de la profondeur [8],
- des épaules : selon les trois axes.

Nous avons adopté la structure d'annotation présentée par la formule ci-dessous(3.1) :

$$(sens)\ mouvement\ ([angle\ /]\ axe) \tag{3.1}$$

- *sens* : Sens du mouvement représenté par un symbole. Vers la gauche ou vers le bas '-', vers la droite ou vers le haut '+'.
- *mouvement* : Pour une rotation marquée, on note 'rot' ; un léger déplacement est noté 'tr'.
- *angle* : symbolise l'angle de la rotation (en degré), l'estimation de l'angle varie de '3' (très faible), à '45' (rotation importante).
- *axe* : symbolise l'axe par rapport auquel s'effectue la rotation, la verticale 'y', l'horizontale 'x', la profondeur 'z'.

8. La description de l'orientation des axes considère une vue de face

Une valeur de rotation peut être combinée de plusieurs types (plusieurs mouvements, axes), dans ce cas, les valeurs sont séparées par des '-'.

La figure (3.4) illustre un exemple d'annotation de mouvements combinés de la tête composés d'un léger déplacement vers la droite et d'une faible rotation vers la gauche selon la verticale '(+) tr (y) - (-) rot (20/y)'. Ces mouvements correspondent en partie à la réalisation du signe [LIEU] illustrés dans les figures (3.4).(2)..(4).

La figure (3.5) illustre un exemple d'annotation de mouvement de déplacement de l'épaule gauche vers le haut '(+) tr (y)'.

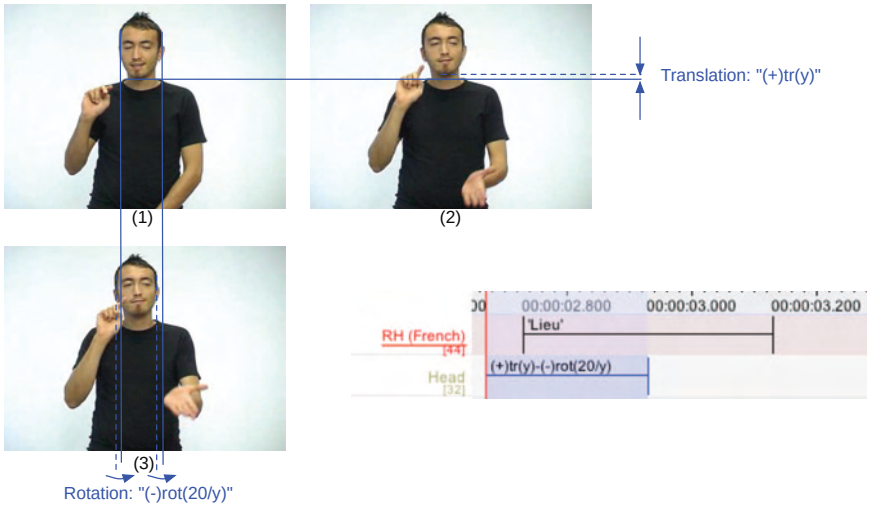


FIGURE *3.4 – Annotation des mouvements combinés de la tête*

Annotation de la cible du regard : Les valeurs des cibles du regard sont :

- La caméra : 'c'
- Une partie du corps du signeur : on y trouve la main droite 'rh' et la main gauche 'lh'.
- L'espace : il a été découpé en trois parties comme l'illustre la figure (3.6).

La délimitation de la valeur de la direction du regard est liée à la stabilisation de la position de l'iris. Le début et la fin correspondent à l'immobilisation de l'iris juste après et avant un changement de position.

L'annotation de la cible du regard nécessite, en plus, de comprendre la structure du discours (les entités, leur emplacements, etc.), une visualisation de la vidéo image par image pour vérifier le début et la fin de changement de la direction du regard ou une directions du regard non perçue pour plusieurs raisons (ex. la signeur baisse la tête) et par conséquent, la cible ne peut pas être déterminée. Nous avons obtenu un taux important de regard perçu (de l'ordre de 82%) par rapport à la durée totale de données annotées (8 vidéos).

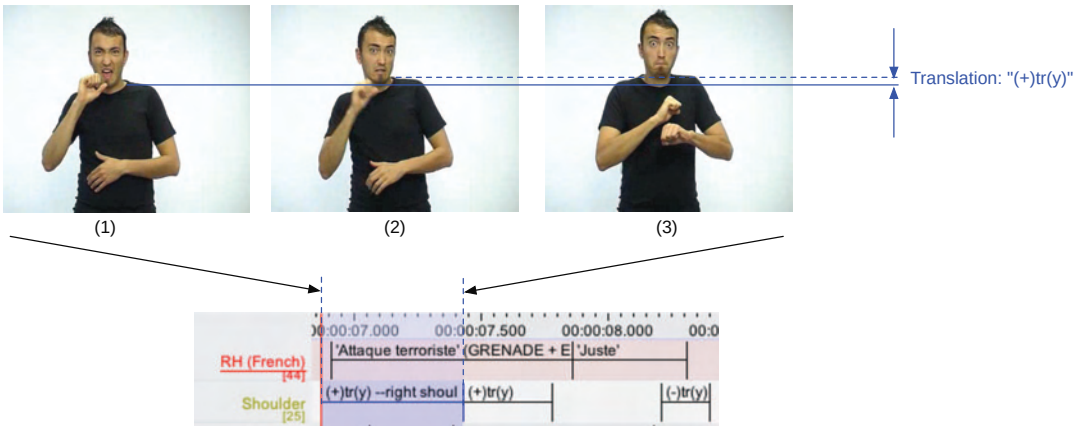


FIGURE *3.5 – Annotation du déplacement de l'épaule gauche vers le haut*

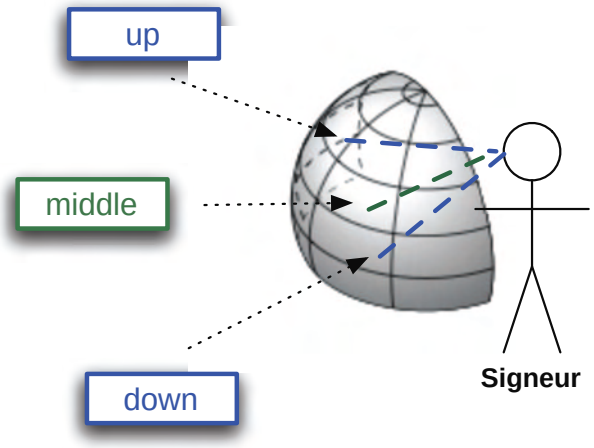


FIGURE *3.6 – Découpage de l'espace de signation en zones ciblées par le regard*

Nous avons rajouté les termes : 'middle', 'up' et 'down' pour différencier les zones spatiales. La figure (3.7) illustre un exemple d'annotation des cibles du regard qui concernent l'espace de signation selon le découpage proposé dans (3.6) ; la zone gauche de l'espace. La cible du regard concerne également la caméra et la main droite : 'left –middle', 'c', 'rh'.

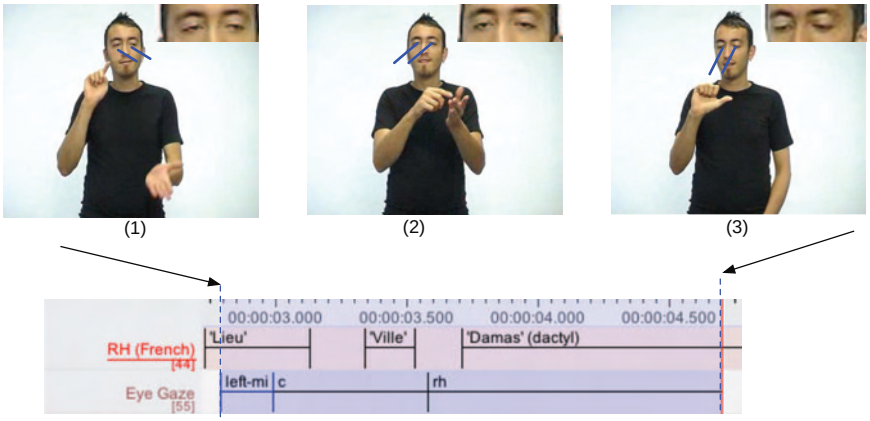


FIGURE *3.7 – Annotation des directions du regard*

Annotation du référencement : L'annotateur repère les moments où le signeur associe un signe à une zone de l'espace. Les valeurs sont marquées dans des pistes nommées « *Entité x - creation* », où *x* représente le numéro attribué à l'entité créée. La figure (3.8.1) est un exemple de création d'un signe [LIEU]. Le résultat est une entité qu'on nomme 'E1'.

Par la suite, l'annotateur marque les instants où le signeur fait référence à ces zones. Les indices visuels de marquage spatial sont les configurations manuelles de pointé, une brève fixation du regard et/ou une rotation des épaules.

Les marqueurs de début et de fin de référencement se manifestent par le changement qui se produit dans l'un des indices suivants :

- Le début d'un signe de pointé est marqué par la stabilisation de la configuration en question. La fin correspond au premier changement de configuration manuelle perçu.
- Le début et la fin d'une fixation correspondent au déplacement de l'iris qui précède et qui suit la fixation.
- Le début de balancement du buste correspond à la rupture avec une posture fixe. La fin du balancement se traduit par une posture fixe.

La figure (3.8.2) est un exemple de référencement d'une zone de l'espace allouée au signe [LIEU]. La valeur 'E1' est le nom de l'entité désignée.

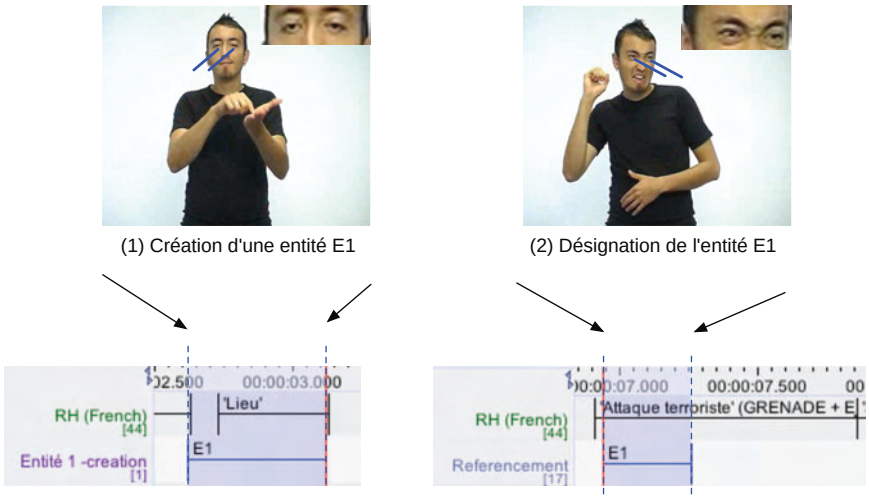


FIGURE *3.8 – Construction et référencement d'un entité*

Annotation du type de référencement : Dans le but de faire apparaître des classes de référencement, nous avons rajouté une piste *« Type de référencement »* où l'annotateur mentionne le type de combinaison gestuelle du référencement en fonction de la présence ou non de valeurs dans les pistes :

- Main droite (1)
- Direction du regard (2)
- Epaules (3)
- Tête (4)

Les valeurs possibles du type de référencement sont :

'M' : Signes réalisés par la ou les deux mains seulement.

'NM' : Signes réalisés par au moins un composant parmi cette liste {la tête, le buste et le regard}.

'MNM' : Signes réalisés par la ou les deux mains et au moins un composant parmi cette liste {la tête, le buste et le regard}.

L'attribution des valeurs est régie selon les règles suivantes :
La main seule :

$$(1) \wedge \neg(2) \wedge \neg(3) \wedge \neg(4) \Rightarrow \text{'M'} \tag{3.2}$$

Tout sauf la main :

$$\neg(1) \wedge (2;3;4) \Rightarrow \text{'NM'} \tag{3.3}$$

La main et au moins un des autres composants :

$$(1) \wedge (2;3;4) \Rightarrow \text{'MNM'} \tag{3.4}$$

Notons que les valeurs 'M', 'NM' et 'MNM' ont été attribuées manuellement. (2 ;3 ;4) : Toute combinaison temporelle impliquant les composantes corporelles (2), (3) et (4).

Par la suite, nous avons explicité les valeurs 'NM' en mentionnant les combinaisons gestuelles du référencement telles qu'elles sont réalisées. Le figure (3.9) illustre d'une part un exemple d'annotation des mouvements non-manuels et d'autre part la délimitation d'une séquence de référencements. La piste *« Combinaison »* permet d'identifier les composantes corporelles qui réalisent le référencement délimité dans la piste *« Type de référencement »*. Par exemple, la combinaison 'TR' (tête puis regard) identifie les composantes corporelles réalisant le référencement de type [NM] (Voir figure 3.9).

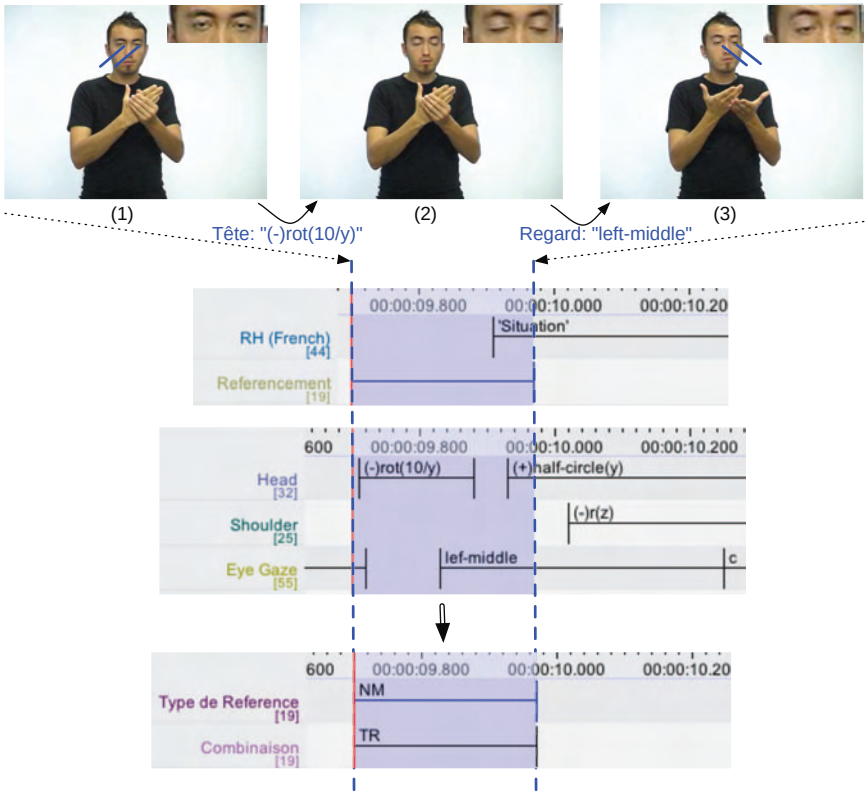


FIGURE *3.9 – Exemple d'identification des composantes corporelles réalisant un référencement de type [NM].*

Déroulement de l'annotation : Les annotations réalisées manuellement sur les corpus de modélisation (Sign'Com 1, Marqspat et Websourd) regroupent d'une part les fonctions de référencement, de localisation de signe et d'autre part la forme du mouvement manuel et non-manuel.

Les annotations qualitatives de la forme des gestes ont été réalisées indépendamment de l'annotation de fonctions [9] par le même annotateur sans vérification.

Les annotations de fonctions de référencement et de localisation ainsi que des gloses (signifiant) ont été réalisées par un entendant débutant en (LSF) et vérifiées par un expert en (LSF) sourd de naissance. L'opération de vérification se résume en ces étapes :

1. L'expert prend connaissance du contenu de l'énoncé,
2. Elimine les segments qui ne correspondent pas à un référencement,
3. Réajuste les limites des segments de référencement et rectifie les motifs de gestes correspondant au référencement.

Quelques chiffres : Nous avons annoté 9 vidéos d'une durée totale de 5 minutes et 31 secondes. L'annotation de ces vidéos a été réalisée en 33 heures.

3.2.3.3 Constatations

L'analyse préliminaire des deux corpus nous a permis de déduire que, d'une part, le nombre d'occurrences des référencements manuels est peu élevé et est largement inférieur à celui des référencements non-manuels.

3.3 Corpus d'exploitation

Il s'agit d'un second corpus, en langue des signes française (LSF), réalisé dans le cadre du projet ANR *« SignCom »*. Il se compose de vidéos et de données $3D$ de capture du regard. Le thème du corpus est la nourriture, en particulier, les recettes. Le corpus se compose de quatre exercices d'interaction entre un signeur réel et un avatar signant (Voir figure 3.10). Il a été élaboré dans le but d'évaluer le système interactif d'apprentissage de la langue des signes française. Rappelons que l'objectif est d'obtenir une modélisation $2D$ du référencement. Par l'utilisation de ce corpus, nous envisageons vérifier si le modèle d'aspect décrit bien la réalisation gestuelle du référencement. Les enregistrements vidéo représentent la base sur laquelle nous effectuerons des mesures $2D$, puis des interprétations de ces mesures en vue d'enrichir le modèle $2D$. Les données de capture de regard feront également l'objet de mesures et d'interprétations $3D$ en vue d'apporter des éléments de réponse à la consistance de la donnée de la direction du regard [10]. De cela, nous avons mis en place un corpus qui inclut :

- Des vidéos enregistrées dans des conditions qui permettent leur exploitation en termes de traitement d'images.
- Des données de capture du regard.

Dans quatre scénarios, le signeur est appelé à pointer des zones de l'espace de signation partagé entre l'avatar et le signeur. Le premier exercice consiste à réordonner

9. chaque forme correspondante à un mouvement d'une composante corporelle a été annotée dans une ligne part

10. La consistance de la donnée du regard signifie qu'elle suffit pour déterminer l rôle du regard

des images projetées dans le but de construire une histoire cohérente. Dans le second exercice, le signeur est appelé à trouver l'ingrédient manquant dans le but de compléter la liste de signes réalisés par l'avatar en se référant à la liste projetée. Dans le troisième exercice, l'avatar demande au signeur réel de pointer l'endroit où se trouve l'ingrédient nécessaire à une recette donnée. Le dernier exercice consiste à choisir un plat parmi ceux qui sont proposés et de demander de changer les ingrédients ou de modifier les quantités proposées. Il a été décidé de construire ce corpus de *« SignCom »* en l'adaptant pour enrichir les modèles de référencement d'où l'introduction d'exercices supplémentaires.

Nous avons demandé aux participants de décrire deux images projetées à l'écran. La description d'un voyageur en attente dans un arrêt de bus favorise l'utilisation de l'espace de signation et donc les référencements afin de mettre en avant les détails de la scène (banc, personne, bus, etc). La description d'une personne qui tombe dans les escaliers favorise l'utilisation des expressions faciales et des gestes afin de décrire ce que ressente la personne (la personne est effrayée, a mal, etc).

Lors de l'enregistrement, nous avons fait appel à plusieurs signeurs sourds et entendants pratiquant la langue des signes française. L'équipement d'enregistrement se compose d'une caméra focalisée sur le signeur et d'un équipement de capture du regard *« Facelab »*. Le début de chaque session est précédé d'une phase de calibrage composée d'une séquence de signes de pointé [11] et par le regard vers les zones marquées de l'écran afin d'établir la correspondance entre zone projetée et zone pointée réellement.

3.4 Corpus d'évaluation

Etude des profils de vitesses : Dans le but d'évaluer les modèles $2D$ de référencement, nous avons utilisé un corpus vidéo élaboré dans le cadre de l'atelier *« DEGELS »* (DEFI Geste et Langue des Signes). La captation a été réalisée dans la chambre sourde du LPL avec deux caméras mini DV semi pro CANON XM2 et une caméra HD Sony hdr-cx 550 VE pour la vue d'ensemble. http ://degels2012.limsi.fr. Le corpus consiste en un dialogue en (LSF) portant sur une description détaillée des lieux à visiter à Marseille : le vieux port, l'église, le château d'If, etc. Ainsi, la vidéo comporte plusieurs signes de pointé déployés par le signeur pour bien marquer les zones localisées et éviter une description ambigüe.

Etude des distances composante corporelle – locus : Nous avons réutilisé une partie des vidéos réalisées dans le corpus *« SignCom »* décrit dans (3.3). Les vidéos en question sont celles de description d'images de *la station de bus* et *de l'homme qui tombe*. Les conditions d'enregistrement justifient notre choix de ce corpus. Elles consistent en un fond uniforme et le port d'habits sombre à manches longues pour faciliter le repérage de

11. réalisés dans la plus part des cas par la main droite. Nous avons aussi observé des signes de pointé réalisés par la main gauche de signeur droitier

FIGURE *3.10 – Corpus « SignCom »*

la tête et des mains.

Aussi, parce que les images projetées sollicitent une description structurée :

– *Station de bus* : le lieu, la personne assise et le bus en stationnement.
– *l'homme qui tombe* : les escaliers, les affaires posées par terre et l'homme qui glisse.

et donc l'utilisation de l'espace de signation pour localiser des signes, indiquer leur proximité et pour y faire référence.

3.5 Conclusion

Nous avons présenté le cahier des charges des corpus qui permettent de construire des modèles décrivant les gestes de référencement et par la suite de les évaluer par un système de détection automatique de référencement. Compte tenu des corpus dont nous disposons, nous avons choisi d'associer à chaque étape (modélisation, exploitation et évaluation) un corpus différent. Toutefois, les corpus choisis ne remplissent pas nécessairement toutes les conditions imposées par le cahier des charges.

Modélisation et exploitation : Le corpus de modélisation présente des taux de perte des données $3D$ de capture du regard. Le tableau (3.2) résume les taux de perte estimés dans les données du corpus *« Marqspat »*. Les données de capture de mouvement présentent des confusions entre marqueurs. Cela a nécessité de mettre en place des protocoles de correction de données et de remplissage des trous par interpolation dans la mesure du possible. Le protocole de remplissage de trous consiste à estimer les valeurs absentes par le calcul de la moyenne mobile des valeurs de voisinage sur des données perdues dont la durée ne dépasse pas 250 ms [12].

Participant	**Taux d'occultation (%)**
Sourd1	21.2
Sourd2	26.45
Entendant1	9
Entendant2	60
Entendant3	65.8
Entendant4	71.3

TABLE *3.2 – Répartition des taux de perte de données du regard par participant*

Nous avons besoin d'interpréter l'association données $3D$ de capture de mouvement et du regard qui présentaient des décalages temporels et n'étaient pas exprimées sur la

12. la durée moyenne d'un clignement

même échelle. Nous avons donc été amenés à automatiser la synchronisation temporelle et spatiale des données de capture de mouvement et du regard.

L'annotation des corpus de modélisation et d'exploitation est manuelle et donc une tâche fastidieuse. Le protocole d'annotation décrit dans la sous-section (3.2.3.1) n'a pas été donc respecté dans toutes les vidéos ce qui a nécessité une révision complète de l'annotation. Vu le temps considérable de révision de l'annotation, nous avons utilisé les données annotées qu'elles soient vérifiées ou pas.

Evaluation : Le corpus *« DEGELS »* consiste en une vidéo de dialogue. Nous avons observé des signes réalisés par les deux personnes en même temps. Afin d'obtenir une segmentation non ambigüe des régions d'intérêt (main et tête) des deux signeurs, nous avons pris le parti de développer une application d'annotation graphique qui permet de récupérer manuellement les coordonnées spatiales des régions d'intérêt (Mains et tête).

CHAPITRE 4

Le référencement : Modélisation

Dans ce chapitre, nous utiliserons les corpus décrits dans le chapitre précédent et analyserons les gestes de référencement afin de déterminer leurs caractéristiques temporelles et spatiales. Nous conclurons par des mesures qui enrichissent le modèle de référencement résultant.

Sommaire

4.1 Objectif

Nous avons annoncé dans la section (2.5) du deuxième chapitre que la méthode de construction d'un système de reconnaissance de gestes de référencement va se baser sur deux axes :

- Les propriétés temporelles des gestes telles que le décalage temporel entre gestes,
- Les propriétés spatiales telle que le comportement des gestes par rapport à une même zone de l'espace (le locus).

Les propriétés temporelles des gestes sont l'ordre dans lequel sont réalisés les gestes de référencement et le décalage temporel entre eux. Tous deux apportent des informations utiles à la modélisation du référencement tel qu'il a été réalisé par le signeur et utile à la reconnaissance d'une séquence de référencement à partir d'un flux vidéo.

La modélisation : Le résultat de la modélisation du référencement consiste donc à construire : 1) un modèle temporel qui tient compte de la chronologie et du décalage temporel entre gestes et 2) un modèle spatial qui exprime la position spatiale des composantes corporelles et linguistiques qui jouent un rôle dans le référencement. Le modèle temporel décrit les combinaisons temporelles possibles dans la réalisation d'un référencement ainsi que le poids correspondant à chaque combinaison. Exemples : la tête, le

regard puis la main ; Le regard, la tête, l'épaule puis la main, etc. Le modèle temporel précise la simultanéité des gestes de référencement annoncée par les linguistes en terme de décalage temporel. Exemple : Le regard précède la tête de 3 images [1], le mouvement de la tête dure 7 images et chevauche le mouvement de la main de 5 images.

D'autre part, le modèle spatial précise la variation du décalage spatial entre composantes corporelles jouant un rôle dans le référencement et la zone cible (un ou plusieurs locus).

La reconnaissance : Le modèle temporel représente une information utile à la segmentation de séquences de référencement. D'une part, connaissant les composantes corporelles en mouvement ainsi que les décalages temporels entre les débuts des mouvements (à partir du flux vidéo), nous avons une information a priori sur la nature de la séquence en cours et sur la durée de la séquence de référencement. Afin de réduire le coût de détection de composantes corporelles en mouvement, nous nous proposons de vérifier l'existence de similitude entre combinaisons gestuelles en terme de séquencement gestuel et de décalage temporel. La similitude entre combinaisons gestuelles permettra de déduire à partir d'une combinaison gestuelle de degré [2] n une combinaison de degré $n-1$. Ceci permettra de 1) réduire le coût de détection de zones d'intérêts (la main, la tête, le buste et le regard) dans une séquence vidéo et de 2) s'affranchir de la détection de la composante du regard.

Le modèle spatial fournit une estimation de la position de la zone qui sera référencée ce qui permet de restreindre la détection automatique – en terme d'analyse d'images – d'un référencement en une simple opération de détection de la fin du référencement. Concrètement, la détection d'un référencement revient à dire SI la position de la main droite (en cas de référencement manuel) correspond à la position de la zone référencée OU SI la droite qui porte l'orientation de la tête ou de la direction du regard croise la position de la zone référencée.

Finalement : Nous proposons de construire : 1) un modèle temporel décrivant le séquencement de gestes de référencement, 2) un modèle spatial qui consiste, dans un premier temps, en des représentations géométriques du référencement réalisé, séparément, par la main, le regard et la tête. Dans un deuxième temps, nous cherchons à fusionner ces représentations dans le but de représenter l'aspect multi-linéaire.

Pour cela, deux types d'analyses seront détaillées :

1. En ce qui concerne l'aspect temporel, nous listerons les combinaisons gestuelles de référencement et proposons de déterminer une ou plusieurs caractéristiques temporelles des variantes gestuelles.
2. Nous proposons une mise en correspondance entre les gestes de référencement et

1. Les corpus que nous avons traités ont été acquis à la fréquence de 25 images par seconde. La durée d'une image correspond donc à 0.040 seconde
2. Le degré d'une combinaison est le nombre de composantes gestuelles présentes

le contenu de l'espace de signation afin de déduire des relations spatiales.

Afin de concrétiser les analyses énoncées ci-dessus, nous avons besoin de représenter les composantes corporelles pour pouvoir automatiser l'analyse des données des corpus tri-dimensionnels.

4.2 Représentations géométriques

Nous nous proposons de représenter la main dominante [3], la direction de la tête, la cible du regard et le locus. Avant de passer aux modèles géométriques, nous mettrons au clair la différence entre la notion de main dominante / dominée et la main droite / gauche dans le but de lever l'ambiguïté dans l'appellation de la main qui réalise un geste de référencement.

4.2.1 Main droite ou main dominante ?

Dans le cas d'un signe à une seule main, [Chetelat 2010] précise que la main dominante est la main droite pour un signeur droitier (et gauche pour un signeur gaucher). Quand il s'agit d'un signe à deux mains, deux cas se présentent :

- Un signe où les deux mains réalisent des mouvements symétriques [F. Lefebvre-Albaret 2010]. Dans ce cas, on ne peut distinguer la main dominante de la main dominée.
- Un signe où les deux mains réalisent deux mouvements différents. La main dominée est par convention celle qui bouge le moins et qui représente dans une partie de la réalisation du signe un repère fixe pour la main dominante.

Nous présentons dans le tableau (4.1) le nombre de participants dans les sessions d'enregistrement, classé par langue. 12 sourds parmi 15 sont droitiers. Les sourds qui signent en LSQ sont à la fois droitiers et gauchers. C'est-à dire, ils utilisent la main gauche pour réaliser des signes à une main censés se réaliser avec la main droite. Nous avons mesuré les distances élémentaires parcourues par les mains droite et gauche, la moyenne et l'écart-type des distances des deux mains afin de les comparer. La figure (4.1) illustre les moyennes et écart-types de distances élémentaires parcourues par les deux mains. Ces mesures concernent des intervalles temporels de sessions d'enregistrement en (LSQ) où nous avons observé des gestes manuels. Nous constatons que les distances parcourues par la main gauche varient autant que celles parcourues par la main droite. Ceci appuie les remarques au sujet de la détermination de la main dominante. En d'autres termes, la main droite n'est pas toujours dominante, même quand il s'agit d'un signeur droitier réalisant un signe.

Ce que nous en retenons c'est que la main dominante n'est pas toujours la main droite pour un signeur droitier et gauche pour un signeur gaucher. L'appellation de *« Main dominante »* peut donc être associée à la main droite et à la main gauche dans un même

3. Dans les corpus étudiés, la main dominante est toujours la main droite

Langue	Nombre de participants
LSF	9
LSQ	3
ASL Montréalais	3

TABLE *4.1 – Répartition des participants sourds*

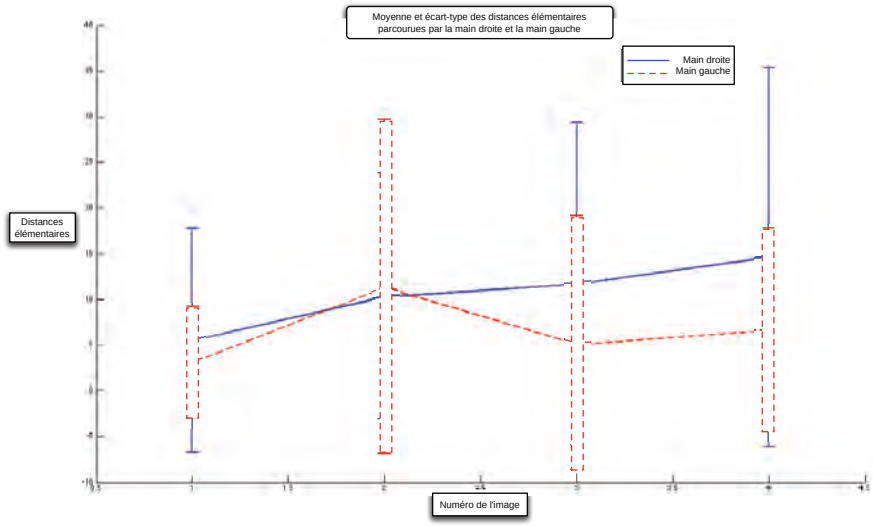


FIGURE *4.1 – Moyenne et écart-type des distances élémentaires parcourues par la main droite et la main gauche des signeurs droitiers*

énoncé. Pour enlever cette ambigüité, nous avons choisi de garder l'appellation *« Main droite »*. De ce fait, l'appellation de *« main droite »* est maintenue dans la suite des représentations et des mesures géométriques.

4.2.2 La main droite

Nous avons exploité les données de capture de mouvement pour estimer une enveloppe sphérique de la main. Les coordonnées $\{x, y, z\}$ du centre C de la sphère sont calculées selon la formule de centre de gravité (4.1).

$$C_{x,y,z} = \frac{RULN_{x,y,z} + RRAD_{x,y,z} + RIND_{x,y,z} + RPIN_{x,y,z}}{4} \tag{4.1}$$

$RULN_{x,y,z}$, $RRAD_{x,y,z}$, $RIND_{x,y,z}$ et $RPIN_{x,y,z}$: représentent les coordonnées des marqueurs de la main droite illustrés dans la figure (4.2)

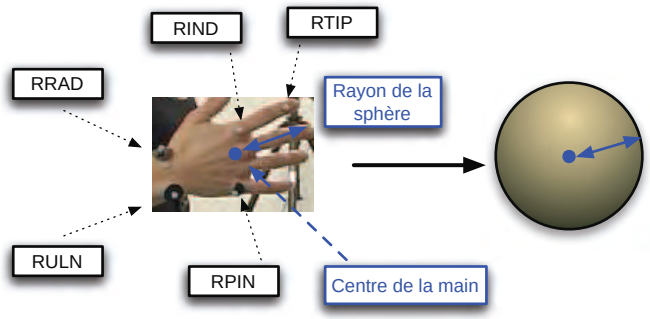


FIGURE *4.2 – Représentation géométrique de la main*

4.2.3 L'orientation de la tête

L'orientation de la tête est mesurée comme étant la normale à la droite passant par les marqueurs *RHEA* et *LHEA* et passant par le marqueur *FORH* comme illustré dans la figure (4.3). Le vecteur $\vec{n}$ est normal au plan formé par les marqueurs : *RHEA*, *LHEA* et *TOPH*. La figure (4.3) illustre le plan en question en vue de dessus. Les coefficients du vecteur normal sont calculés selon le produit vectoriel (4.2.3).

$$\vec{n} = \vec{AB} \wedge \vec{AC} = \begin{pmatrix} TOPH_x - RHEA_x \\ TOPH_y - RHEA_y \\ TOPH_z - RHEA_z \end{pmatrix} \wedge \begin{pmatrix} TOPH_x - LHEA_x \\ TOPH_y - LHEA_y \\ TOPH_z - LHEA_z \end{pmatrix}$$

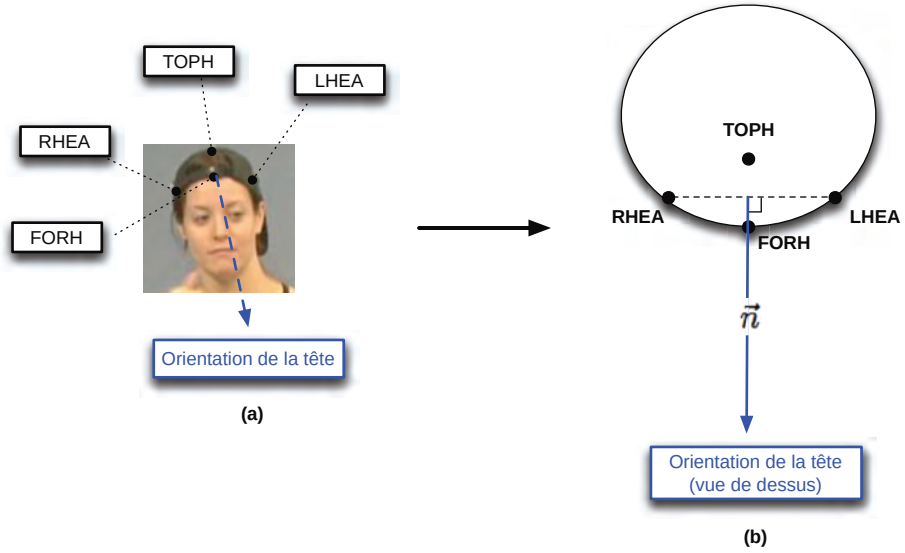


FIGURE *4.3 – a) Marqueurs de la tête, b) Représentation géométrique de l'orientation de la tête (Vue de dessus)*

4.2.4 La cible du regard

La cible du regard représente une reconstruction spatiale de la position ciblée par le signeur à un instant donné. Elle est représentée par un point dont les coordonnées sont fournies par le système de suivi du regard *« FaceLab »*. Nous avons récupéré, à chaque instant, les mesures suivantes :

1. Les coordonnées cartésiennes de la cible du regard.
2. La distance inter-occulaire du signeur.
3. La distance de vergence séparant le milieu inter-occulaire et la cible du regard.

4.2.5 Le locus

Le locus représente une reconstruction spatiale de l'évènement d'assignation d'un signe à une zone spatiale. Les coordonnées spatiales du locus sont, ainsi, liées à celles de la main droite en phase de construction de signe. Nous avons choisi de représenter le locus par l'union des sphères représentatives de la main dominante à chaque instant de la localisation du signe.

Reconstruction du locus : La figure (4.4) illustre la localisation d'une [CHAISE] dans l'espace de signation. La méthode de reconstruction du locus suit les étapes suivantes :

1. la représentation géométrique de la main (Voir 4.2.2),

2. l'enregistrement et les mesures des coordonnées de la représentation géométrique de la main à chaque instant de localisation. Les images de la figure ci-dessous représentent les images clés de la localisation,
3. l'ensemble des coordonnées représente la délimitation du locus dans l'espace de signation.

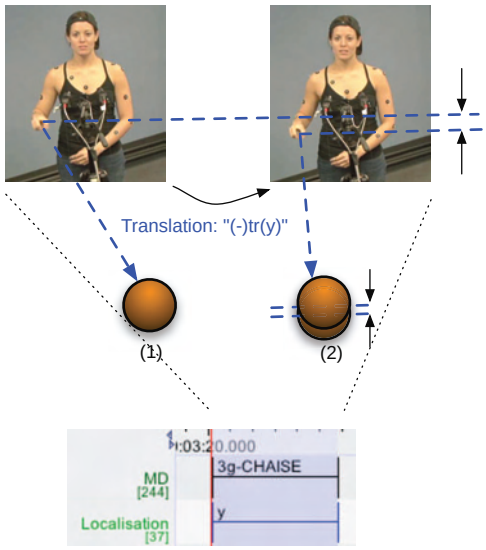


FIGURE *4.4 – Illustration d'une représentation géométrique simplifiée de la zone spatiale assignée au signe [CHAISE] (Vue de face)*

4.3 Paramètres du modèle

Les spécifications de la modélisation géométrique des composantes corporelles et linguistiques étant posées, comment maintenant décrire le geste ou les gestes de référencement en mettant en jeu ces représentations géométriques ? Dans cette partie, nous nous proposons de décrire trois caractéristiques gestuelles de référencement : 1) La dynamique du geste de la main réalisant un signe de pointé, 2) les propriétés temporelles des combinaisons gestuelles, 3) les relations spatiales qui lient les composantes corporelles au locus (ou loci : groupe de locus).

4.3.1 La dynamique du geste de référencement

4.3.1.1 Le pointé

Le geste de pointé manuel ou du *« pointé »* est universel.

Il est déployé dans la communication en langues orales et signées. Cependant, ce type de geste a été peu abordé dans les travaux de traitement automatique des langues (TAL). Nous avons mentionné que ce geste dépend de l'emplacement de départ de la main ainsi que de la position du locus dans l'espace de signation ce qui exclut la possibilité de décrire le geste de pointé en utilisant toutes les caractéristiques d'un signe (l'emplacement, le mouvement, la configuration et l'orientation). En langue des signes française, [Dalle 2009] a observé un comportement caractéristique du pointage isolé. Il décrit l'évolution de l'abscisse instantané de la main au cours d'un pointage. Le comportement présente trois phases :

- un temps d'arrêt,
- un mouvement balistique,
- un temps de pause "relativement long".

Le profil correspondant à ce modèle descriptif est illustré dans (4.5).

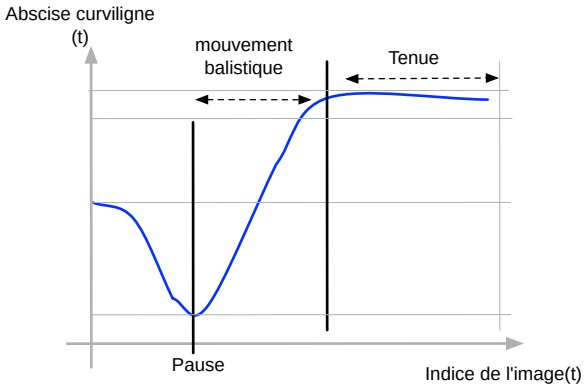


FIGURE *4.5 – Evolution de la vitesse instantannée du pointage*

Nous nous proposons dans un premier temps de quantifier l'amplitude des vitesses instantanées de chaque phase. Puis, nous exploiterons le profil de vitesse résultant afin de retrouver des motifs correspondants sur des mesures effectuées à partir des corpus vidéo décrit dans le chapitre(3).

Nous mesurerons la norme de la vitesse instantanée $\| V \|$ (4.2).

$$\| V \| = \sqrt{V_x^2 + V_y^2 + V_z^2} \tag{4.2}$$

V_x : représente la dérivée du déplacement de la main droite selon l'axe des abscisses (4.3).

$$V_x = \frac{\partial C_x}{\partial t} \tag{4.3}$$

C_x : représente l'abscisse du centre de gravité de la sphère (voir 4.1).

4.3.1.2 La dynamique de la rotation de la tête

Dans le but de mesurer la variation de l'orientation de la tête au cours d'un événement de référencement, nous nous proposons de mesurer l'angle formé par deux vecteurs porteurs des droites D_1 et D_2 d'orientation de la tête correspondante à deux instants distincts.

L'angle α est mesuré selon l'équation (4.4).

$$cos(\alpha)\frac{\vec{n_1}.\vec{n_2}}{\sqrt{\| \vec{n_1} \| . \| \vec{n_2} \|}} \tag{4.4}$$

$\vec{n_1}$ et $\vec{n_2}$ sont les vecteurs directeurs de D_1 et D_2 respectivement dont les coefficients sont calculés selon le produit vectoriel (4.2.3).

4.3.2 Les combinaisons gestuelles

L'analyse multi-linéaire de gestes de référencement repose sur l'aspect temporel. Ce dernier porte sur les décalages temporels entre gestes.

Notations : Afin de faciliter l'extraction de combinaisons ainsi que leur analyse, nous avons eu recours à un système de notation qui simplifie l'appellation de "composante corporelle en mouvement" en une lettre "R", "T", "E" et "M".

« R » : Changement de la direction du regard.

« T » : Mouvement réalisé par la tête (rotation, inclinaison, etc).

« E » : Mouvement de l'épaule (rotation, balancement, etc).

« M » : Mouvement de la main droite réalisant a) un signe localisé ou b) un signe de pointé vers un ou plusieurs loci [4].

Nous avons mesuré les délais temporels entre les gestes de référencement. Pour cela, nous avons utilisé les annotations manuelles effectuées sur le corpus de *« Websourd »*, extrait les débuts et fins des segments appartenant aux pistes : *« RH »*, *« Head »*, *« Shoulders »* et *« Gaze »*.

Outre la possibilité de décrire les intervalles de combinaisons gestuelles, Nous nous proposons d'établir des relations temporelles entre gestes. Un travail similaire a été élaboré dans le cadre de l'étude de la liaison regard – geste en (LSF). [Papazoglou 2010] a utilisé la logique temporelle d'Allen [Allen 1994] et a conclut que les composantes corporelles se comportent d'une manière variée dans leurs relations avec le regard. D'autre part, son étude a révélé que le décalage temporel regard – pointé est de 8 images similaire à celui de regard – signe standard dont la moyenne est portée à 9 images. Nous avons donc utilisé la logique temporelle d'Allen dans le but de :

1. représenter les relations temporelles entre gestes simultanés qui ne commencent pas nécessairement au même moment,

4. La glose correspondante aura pour valeur [PT]

2. établir des relations temporelles directes à partir de relations temporelles indirectes.

La logique d'Allen consiste en 7 classes de relations temporelles ('<','=','m','o','d','s','f') et leurs inverses ('>','=','mi','oi','di','si','fi') [Allen 1983]. On peut illustrer les relations entre intervalles temporels de combinaisons gestuelles en utilisant ces classes. Le tableau (4.2) illustre les relations temporelles possibles entre qui lient le changement de la direction du regard et le changement d'orientation de la tête (dans un référencement).

$'<'$: le changement de la direction du regard (R) est réalisé avant le début de changement d'orientation de la tête (T). Les segments (R) et (T) représentent la durée de réalisation des deux évènements. La relation entre les deux évènements est nommée *« Before(R,T) »* et représentée graphiquement par une "application" d'un noeud N_R (qui représente l'évènement (R)) vers un noeud N_T (qui représente l'évènement (R)). L'application est étiquetée $'<'$ qui symbolise la nature de la relation.

pour les relation $'='$**,** $'m'$**,** $'o'$**,** $'d'$**,** $'s'$ **et** $'f'$: Le segment (R) est délimité par des pointillés verticaux.

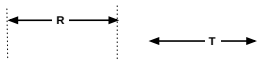
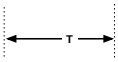
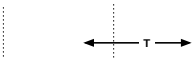
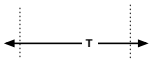
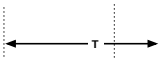
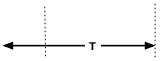
Symbole	Relation	Représentation graphique
$'<' \longleftrightarrow Before(R,T)$		$N_R - (<) \rightarrow N_T$
$'=' \longleftrightarrow Equal(R,T)$		$N_R - (=) \rightarrow N_T$
$'m' \longleftrightarrow Meets(R,T)$		$N_R - (m) \rightarrow N_T$
$'o' \longleftrightarrow Overlaps(R,T)$		$N_R - (o) \rightarrow N_T$
$'d' \longleftrightarrow During(R,T)$		$N_R - (d) \rightarrow N_T$
$'s' \longleftrightarrow Starts(R,T)$		$N_R - (s) \rightarrow N_T$
$'f' \longleftrightarrow Finishes(R,T)$		$N_R - (f) \rightarrow N_T$

TABLE 4.2 – *Formalisme de la logique d'Allen appliqué aux combinaisons de gestes de référencement réalisés à la fois par le regard et la tête (combinaisons de degré 2)*

Les représentations graphiques ne sont valables que pour des relations binaires. Dans la vérité terrain, nous avons identifié des segments de référencement où les évènements

de changement de direction du regard et de l'orientation de la tête se produisent deux fois dans un même segment. Le tableau (4.3) illustre des exemples de cas où l'orientation de la tête change deux fois dans un segment de référencement. La colonne de gauche représente les annotations de changement de la direction du regard (piste *Eye Gaze*) et de l'orientation de la tête (piste *Head*) ainsi que la délimitation des segments de référencement. Nous retiendrons la liste des relations temporelles résultantes dans la même représentation graphique. Exemple : $N_R - (o <) \rightarrow N_T$.

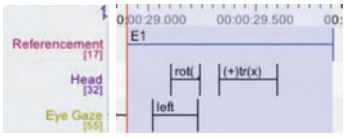
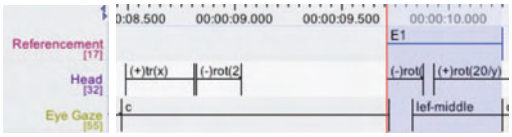
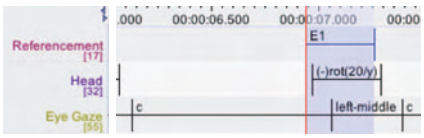
Relation temporelle	**Représentation graphique**
	$N_R - (o <) \rightarrow N_T$
	$N_R - (m < o) \rightarrow N_T$
	$N_R - (o <>) \rightarrow N_T$

TABLE *4.3 – Exemple de formalisation de relations temporelles entre le regard (noté "Eye Gaze" à gauche et "N_R" à droite) et la tête (noté "Head" à gauche et "N_T" à droite)*

D'autre part, la logique d'Allen permet de donner des informations a priori sur la relation temporelle entre événements non consécutifs moyennant la transitivité.

La transitivité : Il s'agit d'un outil de déduction de la forme (4.5).

$$A \overset{r1}{\Rightarrow} B; B \overset{r2}{\Rightarrow} C; \Rightarrow A \overset{r1,r2,r3,\ r4,\ ...,rn}{\Longrightarrow} C \tag{4.5}$$

$r1$: Relation temporelle entre A et B.

$r2$: Relation temporelle entre B et C.

$r1, r2, r3, r4, ..., rn$: Relations temporelles possibles déduites de la matrice de transitivité. $r1$ et $r2$ peuvent en faire partie.

Nous exploiterons cet outil pour deux objectifs :

– Réduire le degré de combinaisons gestuelles comme nous l'avons annoncé dans la section (4.1).

– Réduire les combinaisons gestuelles où le regard est impliqué. Comme nous l'avons expliqué dans le chapitre (3), la direction du regard est un composant

difficile à percevoir dans les conditions actuelles d'acquisition de corpus vidéos et de traitement d'images. Nous essaierons par conséquent d'utiliser la méthode de transitivité dans la logique temporelle d'Allen pour s'affranchir de l'évènement du changement de la direction du regard.

L'exemple illustré dans la figure (4.6) représente une annotation de l'orientation de la tête et du déplacement des épaules dans une séquence de référencement. Les relations d'implications (voir formules ci-dessous 4.6) formalisent certaines relations temporelles sous forme de graphe dont les noeuds $\{N_T, N_R et N_E\}$ représentent l'évènement (le geste d'une composante corporelle) et les arcs $\{oi\ et\ o\}$ représentent le type de relation entre deux gestes.

N_T : la représentation graphique de l'événement gestuel *« Mouvement de la tête »*.

N_R : la représentation graphique de l'événement gestuel *« Changement de la direction du regard »*.

N_E : la représentation graphique de l'événement gestuel *« Mouvement des épaules »*.

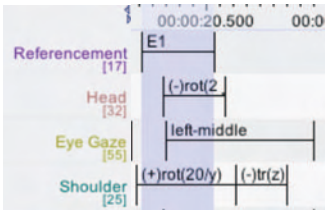


FIGURE *4.6 – Annotation d'une séquence de gestes de référencement*

$$N_T - (o) \rightarrow N_R \qquad (4.6)$$
$$N_R - (oi) \rightarrow N_E$$

D'après le tableau de correspondance établi par [Allen 1983, p. 837], les relations de transitions possibles sont :

$$T(o_{I1I2}, oi_{I2I3}) = (o, oi, d, di) \qquad (4.7)$$

$I1..I3$ sont des intervalles temporels.
o, oi, d et di sont des relations temporelles (Voir tableau 4.2).
Le terme à gauche formalise les deux relations temporelles qui lient, d'une part, $I1$ et $I2$ et d'autre part, $I2$ $I3$. Le terme à droite représente les relations temporelles possibles qui lient $I1$ et $I3$ selon la table de transition. Ainsi, nous avons les relations temporelles

possibles entre gestes éloignés dans le temps du mouvement de la tête et des épaules. Il serait intéressant d'expliciter l'égalité de la formule (4.7) par des coefficients de confiance afin de s'affranchir de la composante du regard.

$$\Rightarrow N_T - (?) \rightarrow N_E$$

Par définition, la transitivité fournit les relations temporelles possibles pour deux événements dont on ne connait que les relations temporelles avec un tiers événement. Il serait intéressant de trouver un outil qui permet de déduire l'inverse c'est-à dire : à partir d'une relation temporelle entre deux événements, on peut déduire deux relations temporelles entre paires d'événements ayant en commun un tiers événement. L'exemple ci-dessous illustre la déduction du changement de la direction du regard (N_R) comme évènement réalisé après le changement d'orientation de la tête (N_R) et avant le mouvement des épaules (N_E). (N_R) est un évènement que nous cherchons à déduire à partir de la relation temporelle directe entre la tête et les épaules formalisée selon la logique d'Allen comme suit : $N_T - (d) \rightarrow N_E$. Concrètement, ce que nous cherchons à déduire ; se sont deux relations temporelles qui lient d'une part le regard et la tête et d'autre part le regard et les épaules. Ainsi, pour un évènement donné (e), ceci serait représenté par la relation d'implication suivante :

$$\exists e \;\; N_T - (d) \rightarrow N_E \overset{?}{\Rightarrow} \begin{cases} N_T - (o) \rightarrow N_R \\ N_R - (oi) \rightarrow N_E \end{cases} \tag{4.8}$$

Il s'agit d'une piste intéressant pour déduire le rôle du regard dans une construction linguistique (référencement par exemple) quand le regard n'est pas perceptible ou difficilement perceptible.

4.3.3 Les relations geste – locus

Nous proposons une étude séparée pour chaque composante corporelle. L'étude de la composante corporelle consiste à l'étude de paramètres mesurables et caractéristiques du geste de référencement selon deux aspects, l'aspect temporel et l'aspect spatial. En ce qui concerne l'aspect spatial, nous avons émis l'hypothèse que la mesure de la position relative de l'entité met en jeu la relation spatiale de la composante corporelle par rapport à celle de l'entité référencée. Pour cela, nous allons étudier les positions relatives composante corporelle – locus.

4.3.3.1 Pointé – locus

Les analyses linguistiques ont mis en avant l'importance de l'accord pointé - regard ainsi que des rôles de l'inclinaison de la tête, du tronc et de la rotation du corps en général dans le marquage de locus que ce soit un marquage spatial unique ou combiné (2.3.2).

Toutefois, nous avons remarqué que le référencement d'une cible est réalisé par une seule composante corporelle à moins que les loci soient situés dans la même zone de l'espace de signation. Dans ce cas on parle de marquage groupé. Cette étude vise à caractériser la relation geste – locus en se basant sur la variation de la distance qui sépare la composante corporelle ou sa direction du locus. Dans ce qui suit, nous adoptons la représentation détaillée dans la sous-section (4.2.5) pour représenter le locus. On se propose de mesurer la distance d entre deux sphères représentatives de la position de la main droite à deux instants différents :

$$d_{C1C2} = \sqrt{(x_{C1} - x_{C2})^2 + (y_{C1} - y_{C2})^2 + (z_{C1} - z_{C2})^2} \tag{4.9}$$

où $C1$ et $C2$ sont les centres des sphères.

La figure (4.7) illustre la distance mesurée partir de deux images clés de localisation du signe [CHAISE].

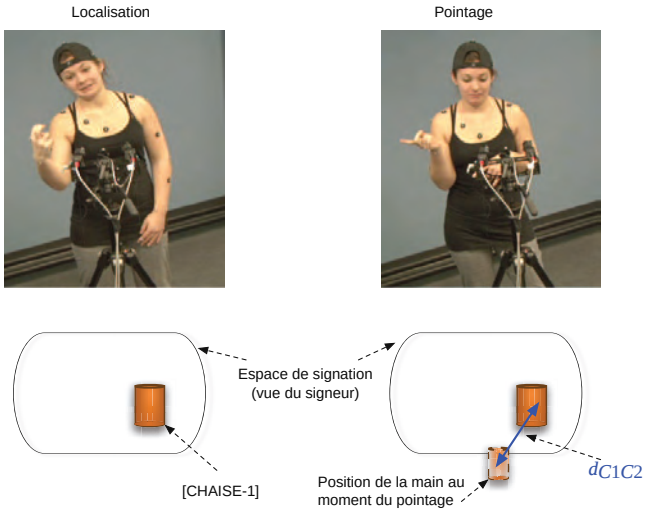


FIGURE *4.7 – La distance mesurée d_{C1C2} à partir de deux images clés de localisation du signe [CHAISE]*

4.3.3.2 Orientation de la tête – locus

On se propose de mesurer la distance d entre la droite portant l'orientation de la tête D et le centre du locus S.

$$d = \frac{\| \vec{n} \wedge S\vec{M}_D \|}{\| \vec{n} \|} \tag{4.10}$$

$\vec{n}$ est le vecteur normal au plan formé par les marqueurs de la tête.

M_D est un point appartenant à la droite D.

La figure (4.8) illustre la distance mesurée entre la droite qui porte l'orientation de la tête et le locus sur une image clé de localisation et de pointage non-manuel du signe [CHAISE].

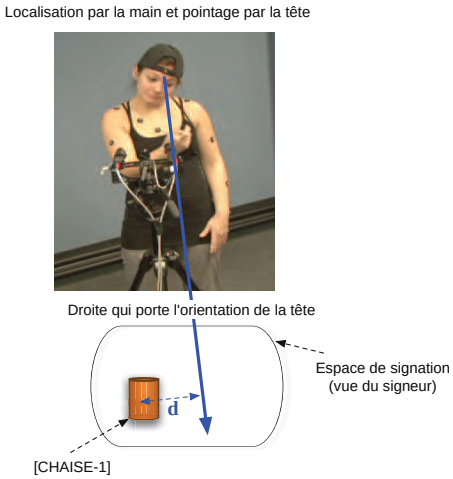


FIGURE *4.8 – La distance mesurée d entre la droite qui porte l'orientation de la tête et le locus sur une image clé de localisation et de pointage non-manuel du signe [CHAISE]*

4.3.3.3 Cible du regard – locus

On se propose de mesurer la distance d entre le point qui représente la cible fixée par le regard et P et le centre du locus S.

$$d = \sqrt{(x_C - x_p)^2 + (y_C - y_p)^2 + (z_C - z_p)^2} \tag{4.11}$$

C est le centre de la main dominante en phase de création d'entité donc C est le centre du locus. p le point cible du regard.

La figure (4.9) illustre la distance mesurée entre la position du point de vergence et celle du locus sur une image clé de localisation et de pointage non-manuel du signe [CHAISE].

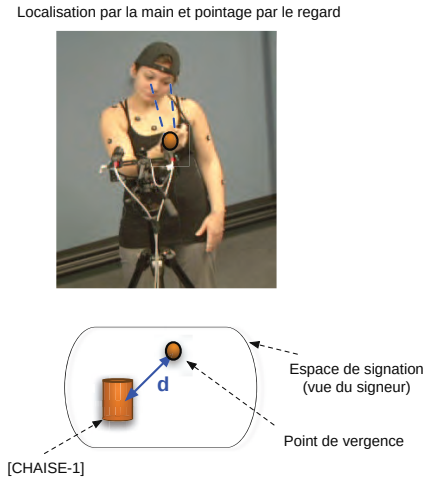


FIGURE *4.9 – La distance mesurée d entre la position du point de vergence et celle du locus sur une image clé de localisation et de pointage non-manuel du signe [CHAISE]*

4.4 Quantification des paramètres

Dans cette partie, nous présenterons les classes de combinaisons temporelles de référencement extraites du corpus *« Websourd »* en utilisant la logique temporelle d'Allen. Nous présenterons également les mesures auxquelles nous avons abouti en utilisant les modèles géométriques proposés appliqués aux mesures de capture de mouvement et du regard réalisées sur le corpus *« Marqspat »*.

4.4.1 Analyses temporelles

Parmi les 9 vidéos de la série (*S* 1) du corpus *« Websourd »*, nous avons recensé en total 55 séquences de référencement dont les taux sont classés par type et par combinaison.

4.4.1.1 Types de référencement

Les types de référencement extraits de la piste *« Type de référencement »* ('M', 'NM' et 'MNM') selon le schéma d'annotation détaillé dans la figure (3.9 :3) sont listés dans le tableau (4.4).

Nous avons mesuré l'évolution de la durée des référencements ('M', 'NM' et 'MNM') et trouvé que les durées du référencement ('MNM') sont largement supérieures à celles de référencements ('M' et 'NM') (Voir figure 4.10).

Type	Taux (%)
'M'	9
'NM'	16.4
'MNM'	74.5

TABLE *4.4 – Taux de fréquence du référencement classés par type, le nombre total de références est 55 dont 6 répétitives de type 'MNM-Ei'*

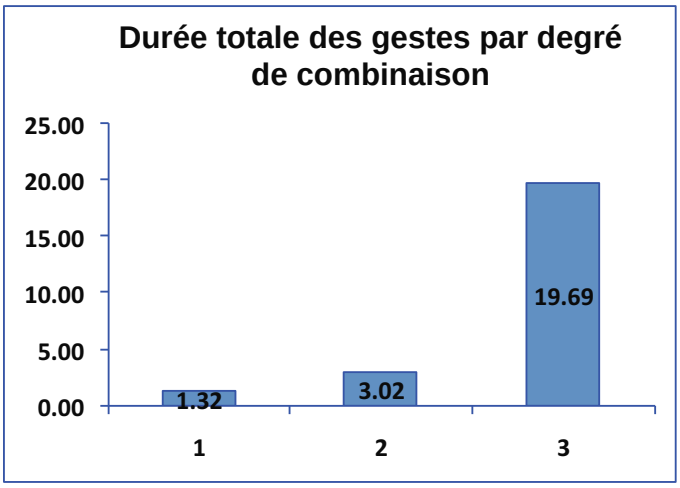


FIGURE *4.10 – Durée des séquences de référencement selon le type de combinaison. 1) 'M', 2) 'NM' et 3) 'MNM'*

4.4.1.2 Combinaisons n-aires

Les combinaisons de gestes de référencement se déclinent en quatre classes : quaternaires, tertiaires, binaires et un-aires. Les motifs extraits de la piste *« Combinaison »* selon le schéma d'annotation détaillé dans la figure (3.9 :3) sont listés dans le tableau (4.5)

Le nombre de combinaisons de gestes sont au nombre de 49. Nous notons que les références vers plusieurs locus sont étiquetés 'MNM-Ei'. Les références correspondants à ce motif 'MNM-Ei' nous amène à nous interroger sur la correspondance : geste – cible. En d'autres termes, nous serons amenés à identifier deux variables : quels gestes pointent vers quelle cible ? Afin d'éliminer cette complexité, nous avons exclu de notre analyse les références réalisés vers plusieurs cibles. Nous traiterons ce type de référencement dans la sous-section (4.4.2) consacrée à l'analyse spatiale.

Nous retenons que les motifs ETRM et TM sont les plus productifs parmi les motifs de combinaisons gestuelles de degrés 4 et 2, respectivement, avec les taux respectifs

Type	Combinaisons	Taux (≈ %)
4-aires	{EMRT ;ETRM ;MERT ;METR ;MRET}	31
	{RMTE ;RTEM ;TERM ;TREM}	
3-aires	{EMT ;ETM ;ETR ;MRT ;MTE}	35
	{RMT ;TRM ;TEM ;TER ;TME ;TRE}	
2-aires	{EM ;RM ;TE ;TM}	23
1-aire	{E ;M}	13

TABLE *4.5 – Taux des combinaisons extraites du corpus « Websourd »*

de 33.3% et 36.36%. En ce qui concerne la composition des motifs, nous avons établi l'histogramme (Figure 4.11) qui décrit le nombre d'occurrences de sous-motifs {(EM/ME) ;(ER/RE) ;(ET/TE) ;(MR/RM) ;(MT/TM) ;(RT/TR)} et avons remarqué que le sous-motif (MT/TM) est beaucoup moins présent dans les combinaisons de degré 4.

Les deux constats mentionnés ci-dessus renseignent le comportement des mouvements de la main et de la tête :

– Dans une combinaison de degré 2, ils sont les plus fréquents.
– Dans une combinaison de degré 4, ils ne se succèdent pas.

Ceci permet de restreindre les zones de l'image à analyser et donc d'optimiser le coût de repérage de mouvements. A la suite de repérage de mouvements décalés dans le temps, les constats ci-dessus permettent d'émettre l'hypothèse de la réalisation d'un référencement de degré 4 (La main, la tête, les épaules et le regard) et donc, d'une part, d'orienter le système de détection de mouvement vers la zone des épaules et, d'autre part, de prendre en considération l'hypothèse de référencement de degré 4 afin de s'affranchir de la détection du regard. D'autre part, nous avons mesuré l'évolution de la durée de gestes ('TER', 'ETRM' et 'M', etc.) et trouvé que le graphe (Figure 4.12), qui illustre l'évolution des durées de gestes de référencement en fonction du degré des combinaisons montre un palier dans l'évolution de la durée pour les degrés 1, 2 et 3. Nous avons enregistré une différence moyenne de 0.07 seconde entre un geste unique et une combinaison de deux gestes, 0.02 entre un geste unique et une combinaison de trois gestes. Ces constatations montrent que la durée d'une séquence de référencement pour les combinaisons de degrés 1, 2 et 3 n'est pas proportionnelle aux durées des gestes qui le réalisent. D'autre part, nous avons remarqué un saut de 0.36 seconde dans le passage d'un geste unique à une combinaison de quatre gestes. Cette constatation est pertinente car elle permet d'émettre des hypothèses sur la durée d'une séquence de référencement en fonction du degré de combinaisons gestuelles.

4.4.1.3 Formalisation

Nous avons choisi de représenter les relations temporelles entre événements non consécutifs en utilisant la description de relations binaires décrites par la logique d'Allen

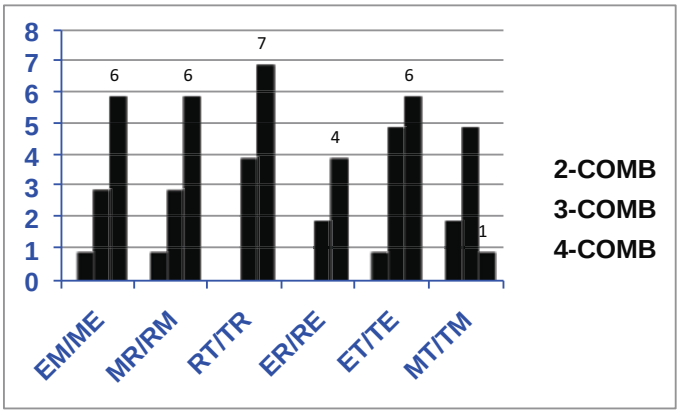


FIGURE *4.11 – Histogramme de nombres d'occurrences de sous-motifs*

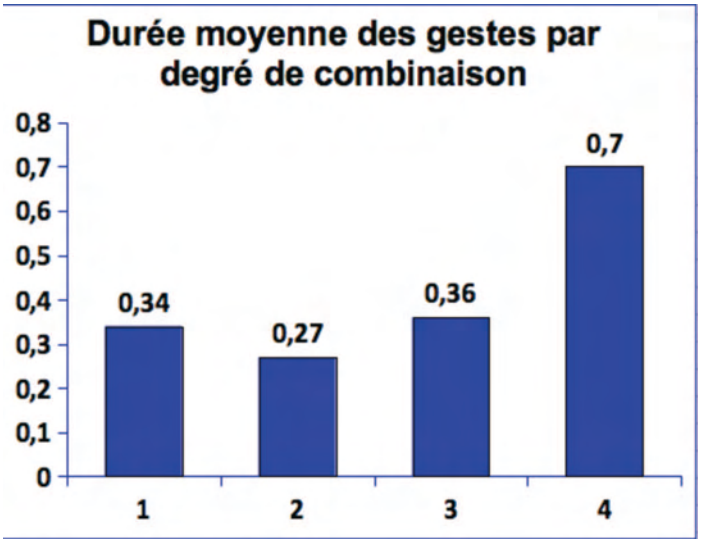


FIGURE *4.12 – Durée moyenne des séquences de référencement selon le degré de combinaison.*

dans le tableau (4.2). Nous nous sommes proposés, également, d'exploiter la matrice de transition pour représenter les relations temporelles entre deux gestes, pas nécessairement consécutifs, ayant un geste intermédiaire. Ceci conditionne donc le choix de la représentation des relations ternaires. En d'autres termes, nous avons choisi de quantifier la relation temporelle qui met en jeu toute paire d'événements qui présente une forme de simultanéité $\{'o';' oi'','m';' mi','s';' f','d';' di'$ ou $' ='\}$. Prenons par exemple le motif 'TRE' qui représente le séquencement mouvement par la tête, changement de la direction du regard puis un mouvement des épaules. Les représentations graphiques possibles sont illustrées dans le tableau (4.6) qui présente les configurations satisfaisant la condition de simultanéité dans les mouvements c'est-à dire que le tableau ne présente que les possibilités où le dernier mouvement (E) chevauche un ou les deux mouvements {T et R}.

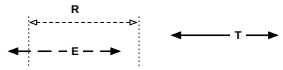
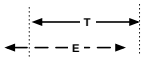
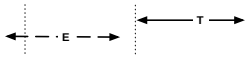
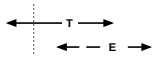
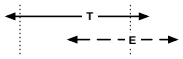
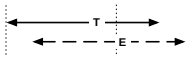
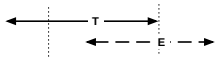
Relations	**Représentation graphique**
	$N_T \leftarrow (<) - N_R - (oi) \rightarrow N_E$
	$N_T \leftarrow (=) - N_R - (oi) \rightarrow N_E$
	$N_T \leftarrow (m) - N_R - (oi) \rightarrow N_E$
	$N_T \leftarrow (oi) - N_R - (<) \rightarrow N_E$
	$N_T \leftarrow (d) - N_R - (o) \rightarrow N_E$
	$N_T \leftarrow (s) - N_R - (o) \rightarrow N_E$
	$N_T \leftarrow (f) - N_R - (o) \rightarrow N_E$

TABLE *4.6 – Exemple de description de relations ternaires basée sur la représentation de la logique d'Allen de la relation temporelle de deux d'événements et sur la condition de simultanéité des événements en question. Le motif 'TRE' représente le séquencement : mouvement par la tête, changement de la direction du regard puis mouvement des épaules.*

La transitivité illustrée par l'exemple (Voir formule 4.7) qui donne les relations temporelles $('o';' oi';' d';' di')$ comme possiblement déduites des deux relations temporelles $('o';' oi')$ – de deux évènements de durées données – nous a permis de conclure qu'on peut obtenir l'ensemble des relations temporelles possibles à partir de deux relations

temporelles ayant un intervalle en commun (voir la liste totale des relations temporelles 4.6).

La transitivité permet de lister des possibilités de relations. Notre objectif final de l'analyse temporelle étant de simplifier les combinaisons de référencement provenant de corpus de modélisation et donc de réduire le nombre de combinaisons possiblement déduites, nous avons recensé les types de relations temporelles entre paires d'évènements dans une combinaison gestuelle (ex. les types de relation entre 'R' et 'T' dans la combinaison RTE. Le calcul de taux de fréquence revient à :

1. Grouper les combinaisons gestuelles similaires de degré 2 (exemple : le groupe RT à n combinaisons).
2. Déterminer les relations temporelles de chaque groupe selon la logique d'Allen (exemple : o_{RT} à m occurrences).
3. Calculer le taux de fréquence de chaque relation (exemple : $\frac{m}{n}$)

Nous avons eu recours à la réduction de degrés de combinaisons moyennant une méthode de déduction (ex. RTEM $\rightarrow$ REM) basée sur la transitivité et des statistiques réalisées sur des combinaisons similaires provenant du corpus *« Websourd »*.

A titre d'exemple, voici les taux de fréquence des sous-motifs EM et MR dans le tableau (4.7).

Sous-motif	**Relation**	**Taux (≈%)**
'EM'	$'o'$	52
	$'di'$	34
'MR'	$'di'$	70
	$'o'$	30

TABLE *4.7 – Taux de fréquence des relations temporelles des motifs* EM *et* MR *dans toutes les combinaisons gestuelles contenant ces motifs.*

Le taux de fréquence d'une relation temporelle représente une hypothèse. En d'autres termes, on peut émettre l'hypothèse que : le motif TE est déduit du motif TRE avec un taux de fréquence du motif déduit TE à partir d'un motif TRE.

Explication : La déduction se base sur la nature des relations temporelles. Une hypothèse sur la relation TE est émise au départ (exemple *o*). Dans le motif TRE nous avons les relations :

– *o* dans TR.
– *oi* dans RE.

Pour le motif TE, l'hypothèse que la relation déduite est de type *d* est possible avec un taux de fréquence de 54%.

Le tableau (4.7) présente les fréquences calculées sur les motifs EM et MR.

L'opération décrite permet de réduire le nombre de combinaisons gestuelles ce qui rejoint notre objectif de la modélisation temporelle. D'une manière générale, le degré d'une combinaison gestuelle est réduit si on peut en déduire un motif existant qui est de degré inférieur. Exemple, le motif EMRT est réduit en EMT si :

$$Rel_{[\mathsf{EMRT}]\mathsf{MT}} = Rel_{[\mathsf{EMT}]\mathsf{MT}}$$

$Rel_{[\mathsf{EMRT}]\mathsf{MT}}$: la relation temporelle qui lie M et T. Il s'agit de la relation déduite moyennant la transitivité d'Allen, exemple : $N_M \leftarrow (d) - N_R - (o) \rightarrow N_T \Rightarrow N_M - (di) \rightarrow N_T$

$Rel_{[\mathsf{EMT}]\mathsf{MT}}$: la relation temporelle qui lie M et T dans ce motif.

Or nous avons :

$$Rel_{[\mathsf{EMRT}]\mathsf{MT}} = di = Rel_{[\mathsf{EMT}]\mathsf{MT}}$$

Donc :

$$\mathsf{EMRT} \Rightarrow \mathsf{EMT} \tag{4.12}$$

De même, nous quantifions les taux de confiance de réductions présentées dans le tableau (4.8). Rappelons que réduire un motif de degré n à un motif de degré inférieur revient à calculer les taux de confiance de l'appartenance de la relation temporelle déduite par transitivité à une des classes de valeurs de relations temporelles mentionnées dans (4.6)

Motif(n)	**Motif**($n-1$)	**Confiance**(%)
EMRT	EMT	100
ETRM	ETM	64
MERT	MRT	50
MRET	MT	50
RTEM	RM	66.7
TERM	TEM	100
MRT	MT	50
TME	TE	100

TABLE *4.8 – Résultats de réductions de motifs de degrés* $n \in \{3;4\}$ *(première colonne) en* $n-1 \in \{2;3\}$ *(seconde colonne) et le taux de confiance correspondant (colonne 3)*

5. Les combinaisons gestuelles de référencement sont au nombre de 43

Type	Combinaisons [5]	Taux de réduction(%)
4-aires	{MERT ;METR}	65
	{RMTE ;TREM}	
3-aires	{EMT ;ETM ;ETR ;MTE}	53.54
	{MRT ;RMT ;TRM ;TEM ;TER ;TRE}	
2-aires	{EM ;RM ;TE ;TM}	80

TABLE *4.9 – Taux de réductions des classes de combinaisons de degrés* 2, 3 *et* 4 *extraites du corpus « Websourd »*

Motif	Taux (%)
ETM	30.43
{ETR ;MRT}	13.04
{MT ;RM ;TE}	25

TABLE *4.10 – Proportions des combinaisons de degrés* 3 *et* 4 *extraites du corpus « Websourd »*

Nous obtenons ainsi une répartition concentrée sur les motifs d'ordres 3 et 2 comme le montre le tableau (4.9). Les proportions de motifs sont présentées dans le tableau (4.10).

Nous nous proposons dans la sous-section suivante de tester ces résultats en appliquant un filtre $F(motif)$ qui décrit les relations ternaires possibles (voir tableau 4.6) pour extraire les combinaisons gestuelles possibles réalisées par les épaules, la tête et le regard {ETR,RTE,TRE,ETR} et les combinaisons {TE et ET} et comparer par la suite les résultats obtenus. Nous avons choisi le filtre F(ETR) car le motif ETR est classé le second motif le plus reproductif après ETM. Nous notons que nous avons exclu les motifs où la main droite est impliquée car la construction d'un filtre comprenant la main droite comme élément requiert la connaissance de la signification du signe réalisé. Les outils dont nous disposons ne permettent pas, pour le moment, de déterminer la configuration de la main droite et donc de la signification du signe produit. De plus, le référencement interprété à partir d'un signe localisé est lié à la signification du signe, ce qui rend l'information fournie par la délimitation temporelle du geste manuel incomplète et donc insuffisante.

Nous avons choisi la notation F(ETR) pour distinguer le filtre de la combinaison ; nous citons comme exemple la combinaison ETR et le filtre F(ETR).

Nous avons utilisé les indicateurs classiques d'évaluation d'algorithmes de classification des langues naturelles. Il s'agit de la précision et de la performance (le rappel) qui

Filtre \ Taux(%)	VP	FP	FN	Precision	Performance	F-measure
ETR	48	0	76	100	38.7	55.8
TE	64	≈37	60	63.3	51.6	56.4

TABLE *4.11 – Résultats de détection de motifs* ETR, TE *parmi* 43 *motifs : 1-3) Taux de VP, FP et FN, 4-6) Interprétations des taux de détection*

signifient respectivement :

- La précision est le nombre de segments de référencement détectés (ou vrais positifs - VP) divisés par le nombre total de segments détectés par le filtre,
- La performance est le nombre de segments de référencement détectés divisés par le nombre total de segments contenus dans l'énoncé.

Nous avons également utilisé la F_{MESURE} (Voir formule 4.15)introduite par [Rijsbergen 1979] qui combine le rappel et la performance permettant ainsi d'obtenir une évaluation exploitable.

4.4.1.4 Détection de motifs 2 et 3-aires

Le tableau (4.11) présente les résultats de détections de motifs de référencement. *VP* : le nombre de motifs de référencement détectés parmi 43 motifs, toutes combinaisons confondues. *FP* : le nombre de faux positifs. *FN* : le nombre de motifs non détectés. A partir des résultats du tableau (4.11- colonnes 1 ;2 et 3), nous avons mesuré la précision et la performance (ou rappel) des résultats de détection (voir formules 4.13 et 4.14), puis combiné les deux indices selon la méthode « F_{MESURE} » (voir formule 4.15) dans le but d'obtenir l'indice d'exactitude qui figure dans le tableau (4.11 - colonne 6) qui montre que les deux filtres permettent de détecter le référencement avec une exactitude quasi-égale.

$$Pre¢ision = \frac{VP}{VP + FP} \tag{4.13}$$

$$Performance = \frac{VP}{VP + FN} \tag{4.14}$$

$$F_{MESURE} = \frac{2 \ . \ Precision \ . \ Performance}{Precision \ + \ Performance} \tag{4.15}$$

Constatations : Parmi les vrais positifs détectés moyennant le filtrage par les caractéristiques des motifs : ETR et TE, c'est-à dire, l'enchaînement des mouvements des épaules 'E', de la tête 'T' puis du regard 'R' pour le motif ETR. Nous avons détecté des motifs autres que ETR et TE tels que : {TRE, RT et ETM}. Ceci nous amène à faire l'hypothèse qu'il existe une relation de la forme :

$$F(\text{ETR}) \overset{?}{\Rightarrow} \{\text{TRE}; \text{RT}; \text{ETM}\} \tag{4.16}$$

Ceci met en avant des relations à plusieurs niveaux mettant en jeu des relations temporelles combinées du type :

$$\text{ETR} \Rightarrow \left.\begin{cases} N_E - (o) \rightarrow N_T \\ N_T - (d) \rightarrow N_R \end{cases}\right\} \Rightarrow N_E - (ods) \rightarrow N_R \leftarrow (oidisi) - N_E \Rightarrow \text{TRE}$$

Les relations temporelles qui lient deux motifs de degrés différents intègrent les relations temporelles combinées ainsi que la suppression et / ou l'ajout d'une ou de deux composantes corporelles comme l'exemple mentionné dans la formule de relation d'implication (4.16). Cela veut dire que les composantes supprimées ou ajoutées (E ; R et M) ne sont pas essentielles pour la détection d'événements de référencement de motifs (RT ; ETM), alors que dans le motif TRE, nous notons, en conséquence, que l'ordre des gestes n'est pas essentiel.

Nous notons ainsi qu'il existe un autre type de relation entre motifs de même classes qui concerne le nombre de composantes corporelles conservées.

D'une manière générale, nous retenons que les motifs de référencement sont variés. Nous avons revisité le classement des combinaisons gestuelles selon le degré de combinaisons en réduisant certaines combinaisons gestuelles de degrés 3 et 4 en 2 et 3 respectivement, moyennant la méthode décrite dans la sous-section (4.4.1.3) de relations temporelles qui lient les motifs. Ce résultat s'avère pertinent pour apporter un élément de réponse à la problématique soulevée dans le paragraphe (4.3.2) et qui concerne l'identification de liens entre motifs. Rappelons que cette problématique est motivée par l'objectif de la substitution de motifs de degré n par des motifs de degré $n-1$. En effet, les liens existants entre mouvements en terme de chronologie temporelle rendent possible la substitution de combinaisons par d'autres selon des contraintes temporelles énoncées précédemment.

4.4.1.5 Interprétations

Nous avons classé les gestes de référencement en trois classes : C1) référencement manuel, C2) référencement non-manuel et 3) référencement manuel et non-manuel.

Les statistiques réalisées sur le corpus *« Websourd »* confirment le fait que les marqueurs non-manuels (C2) sont les plus récurrents dans les discours en langue des signes. Ceci permet d'identifier les éléments à prendre en compte dans la reconnaissance automatique du concept de référencement (Voir tableau 4.12).

La quantification des durées de réalisations gestuelles de chaque type de référencement enrichit la représentation informatique de(s) geste(s) de référencement et apporte de la précision aux modèles linguistiques .En effet, les linguistes expriment la multi-linéarité

Catégorie	Représentation
Geste manuel	Région d'intérêt représentative de la main en mouvement
Geste non-manuel	Région d'intérêt représentative de la tête en mouvement
Geste manuel et non-manuel	Régions d'intérêt représentatives de la main et de la tête, les deux en mouvement avec possibilité de décalage temporel

TABLE *4.12 – Correspondance entre geste(s) de référencement et représentation informatique*

des gestes déployés par une simultanéité. Toutefois, les mesures révèlent un écart marquant dans la durée entre les gestes de référencement manuels 'M' et non-manuel 'NM' et les gestes regroupant les deux, manuels et non-manuels 'MNM'. Ainsi, cette constatation ouvre la voie vers une perspective d'analyse des facteurs provoquant ce décalage et par conséquent permettra discriminer les classes de gestes de référencement ('M' et 'NM') et la classe de référencements à la fois manuels et non-manuels ('MNM').

Les classes de motifs gestuels recensées dans le corpus *« Websourd »* ont montré que les gestes de référencement les plus fréquents sont de type 'MNM', en particulier les gestes de degré 3.

Les relations temporelles entre gestes et motifs gestuels de référencement déduites moyennant la logique d'Allen a fait ressortir plusieurs catégories de relations selon le taux de confiance, le nombre de composantes gestuelles ou encore l'ordre chronologique. Toutefois, nous avons noté qu'il existe un autre type de relation qui met en jeu l'importance ou pas d'un ou de plusieurs composantes corporelles.

4.4.2 Analyses spatiales

Dans cette partie, nous étudierons la distance entre composantes corporelles et loci. Les mesures correspondantes ont été réalisées selon les modèles géométriques présentés dans (4.2). Nous construirons un modèle basé les variations des distances entre composante corporelle et locus. Ensuite, nous vérifierons la validité de ce modèle sur le corpus d'évaluation décrit dans le chapitre (3).

Rappels : Nous avons représenté la main droite par une sphère dont le centre est le milieu de la paume de la main (Voir section 4.2) et le rayon est la longueur de l'index multiplié par $3/2$.

Les sous-sections suivantes présentent le nouveau paramètre introduit ; les distances relatives (Main, regard, tête) – Locus au cours d'un référencement.

4.4.2.1 Modèles de comportement geste - locus

Nous nous proposons dans cette partie de caractériser les gestes de la main droite, de la tête et du regard au cours d'un référencement d'après la variation de la distance qui sépare la position du locus de celle de la composante corporelle. Dans un premier temps, nous évoquerons la méthode de synchronisation de données brutes de capture de mouvement et du regard.

Synchronisation : Compte tenu du volume et de l'hétérogénéité des données dont nous disposons, nous étions amenés à les synchroniser en amont. Nous avons la synchronisation temporelle qui consistait à déterminer le décalage temporel en nombre d'images entre les données fournies par l'équipement de capture de mouvement et celui de *« FaceLab »* sachant que les fréquences d'échantillonnage sont différentes. La méthode se composait de trois étapes [6] :

1. repérer, dans le début de la vidéo, le mouvement le plus saillant comme le haussement de la tête ou de le clap des mains et celui du premier clignement des yeux,
2. repérer, dans le fichier de capture de mouvement, un pic de position spatiale de la tête ou une distance nulle entre les deux mains,
3. repérer, dans le fichier de capture du regard, la première valeur non nulle correspondante au champ *« Clignement des yeux »*,
4. mesurer les décalages entre les instants repères de synchronisations temporelles en tenant compte des fréquences d'échantillonnage de la vidéo, de la capture de mouvement fournie par *« MotionCapture »* et les données de capture de la direction du regard fourni par *« FaceLab »*.

Les mesures de synchronisation temporelle présentent des incertitudes négligeables qui sont dues au passage d'une fréquence à une autre. Les fréquences d'échantillonnage sont de 25 im/sec pour la vidéo et 60 im/sec pour les équipements de capture du regard et des mouvements.

D'autre part, nous avons ramené les coordonnées cartésiennes de la cible du regard, exprimées selon le repère de *« FaceLab »*, au repère du dispositif de capture de mouvement *« MotionCapture »*. Les étapes suivantes résument la méthode de synchronisation spatiale choisie :

1. placer un marqueur de *« MotionCapture »* sur le dispositif de capture du regard,
2. récupérer les coordonnées cartésiennes du marqueur exprimées dans le repère de *« MotionCapture »*,

6. dans le corpus *« Marqspat »*, la synchronisation vidéo – *« FaceLab »* est facilitée grâce à la vidéo intégrée à l'équipement de capture du regard qui renseigne le numéro d'image. Ainsi, il suffit de trouver n'importe quel mouvement commun à tous les équipements

3. soustraire des coordonnées cartésiennes de la cible du regard – exprimées dans le repère de *« FaceLab »* – celles du marqueur posé sur le dispositif *« FaceLab »*.

4.4.2.2 Distances composante corporelle – Locus :

A partir du résultat de synchronisation des données de capture de mouvement et du regard, nous avons appliqué les représentations géométriques illustrées par les modèles géométriques des figures de la table (4.13) sur plusieurs sessions d'enregistrement du corpus *« Marqspat »*.

Rappel : Dans le protocole d'annotation du corpus *« Marqspat »*, un locus est étiquetée par une lettre en minuscule :'*x*', '*y*', '*u*' et '*z*' (Voir exemples dans la figure 4.13). Une zone formée par un groupe de loci est étiquetée par une lettre en majuscule :'*X*', '*Y*', etc. Le protocole d'annotation n'indique pas la proximité des loci groupés. D'autre part, l'exemple d'annotation 'x-y' indique que deux loci sont les cibles successives d'un signe ou d'un mouvement.

Mesures : Les mesures de distances ont été appliquées sur une séquence d'enregistrement d'une durée de 4 minutes et 44 secondes. Nous avons mesuré les distances Composante corporelle – locus telle que la position du locus est la position de la main droite en situation de localisation de zones spatiales selon une lecture transversale des pistes :

1. main droite – locus : *« MD »* et *« Localisation »* (Voir la description de l'annotation dans la section 3.2.3.1),
2. orientation du regard – locus : *« Tête »* et *« Localisation »*,
3. cible du regard – locus : *« Regard »* et *« Localisation »*.

Nous avons sélectionné les séquences de référencement d'un locus {'x', 'y'} et de loci groupés {'x-y', 'y-u' et 'z-y'} et mesuré les distances composante corporelle – locus $D1$. Le tableau (4.14) présente les rapports : $\frac{D1}{R}$ (R : le rayon de la sphère représentative de la main droite) pour :

1. les valeurs minimales des distances mesurées en situation de localisation de loci étiquetés : {*x*, *y*, *x*-*y*, *y*-*u* et *z*-*y*} (Voir figure 4.13),
2. les valeurs minimales des distances mesurées en situation de localisation de 2 locus étiquetés : {*x* et *y*},
3. les valeurs minimales des distances mesurées en situation de localisation d'un locus étiqueté *x*.

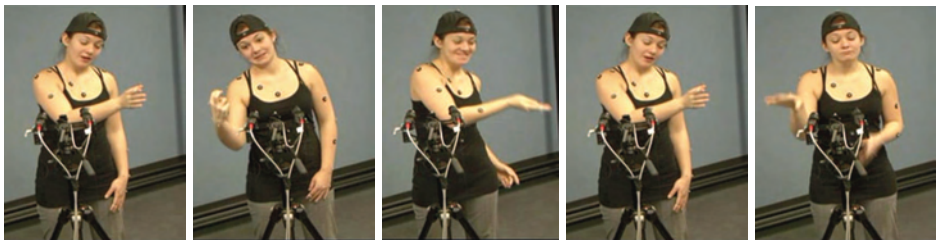


FIGURE *4.13 – Localisations des loci a) x, b) y, c) x-y, d)y-u et e)z-y*

Composante	loci	Taux (%)
Main droite	*x*	**0.33**
	y	1.27
	x-y	1.25
	y-u	1.37
	z-y	**0.06**
Regard	*x*	1.06
	y	2.37
Tête	*x*	**0.01**

TABLE *4.14 – La troisième colonne représente le rapport D/R tels que D : la distance entre les deux centres des sphères représentatives du locus : (x, y) ou loci : (y-u, z-y, x-y)et 1) la main droite, 2) le regard, ou 3) la tête. R : est le rayon de la sphère représentative de la main droite : 104 mm.*

4.4.2.3 Interprétations

Les taux en gras signifient que la composante corporelle réalisant le référencement se positionne à l'intérieur de la sphère représentative de la main droite en phase de création de signe et donc représentative du locus (Voir 4.2). On en déduit que par rapport à la sphère représentative du locus référencé :

- la main droite se positionne à l'intérieur,
- la droite qui porte la direction du mouvement de la tête l'intercepte,
- la position de la cible du regard est à l'extérieur.

Les taux correspondants aux distances Regard – Locus montrent que la distance cible du regard – Locus est inférieure à celle du rayon de la sphère représentative du loci. Par conséquent la direction du regard marque d'une manière précise la zone spatiale occupée par un seul locus.

Nous avons définie la précision d'un référencement par l'intersection entre zone spatiale 3D du locus et représentations géométrique de la composante corporelle 3D. Notre analyse spatiale nous a permis d'identifier les composantes corporelles qui pointent le locus d'une manière précise qui sont la main et la tête.

4.4.3 Analyse de la vitesse

Dans cette partie, nous étudierons deux paramètres du référencement, le profil dynamique de la vitesse de la main droite et la distance entre composantes corporelles et loci. Les mesures correspondantes ont été réalisées selon les modèles géométriques présentés dans (4.2). Nous émettons l'hypothèse que l'évolution de la vitesse de la main droite représente une caractéristique du référencement. Pour montrer l'efficacité et les limites de cette hypothèse, nous construirons un modèle basé sur les durées des profils de la vitesse d'un geste de référencement. Ensuite, nous testerons la validité de ce modèle sur le corpus d'évaluation décrit dans le chapitre (3).

Rappels : La méthode de transcription manuelle des signes de pointé manuel décrits dans le chapitre (3) dépend de deux critères : La stabilisation de la configuration manuelle

 et / ou le mouvement de la main.

4.4.3.1 Profil dynamique de la main droite :

Le but de cette sous-section est de quantifier le profil dynamique de la main droite énoncé par [Dalle 2009] (Voir figure 4.14). Pour cela, nous présentons les mesures de vitesses instantanées tri-dimensionnelle de la main droite obtenues en appliquant la dérivée du déplacement élémentaire par rapport au temps (formule 4.2).

A partir du corpus de capture de mouvement *« Marqspat »* en (LSQ), nous avons trouvé plusieurs profils de vitesse dynamique de segments de pointés manuels délimités selon les critères de transcription. Parmi ces profils, deux signatures caractérisent la majorité des profils dynamiques. La signature $Sign1$ illustrée dans (Figure 4.14) répond au critère du mouvement et peut être à celui de la configuration manuelle – ceci n'a pas été vérifié. Le mouvement commence et se termine avec une vitesse quasi-nulle. La signature $Sign2$ illustrée dans (Figure 4.15) répond au critère de la configuration manuelle uniquement et par conséquent commence par une vitesse non-nulle.

La signature $Sign1$ illustrée dans (Figure 4.14) a été élaborée à partir de 8 pointés manuels. Elle comporte une première phase ↗ et une seconde ↘.

La seconde signature $Sign2$ illustrée dans (Figure 4.15) a été élaborée à partir de 4 pointés manuels. Elle comporte une seule phase ↘. Ce type de signature révèle, certes, un phénomène étrange du fait que le signe correspondant commence à une vitesse non nulle, mais nous tenons à signaler que les référencements analysés (en l'occurrence $S2$ (Figure 4.15) sont le résultat d'une annotation manuelle. Pour délimiter les segments de référencement ou de signes, d'une manière générale, les annotateurs prennent en compte

– C1 : la stabilité de la configuration manuelle et donc de la forme du signe.

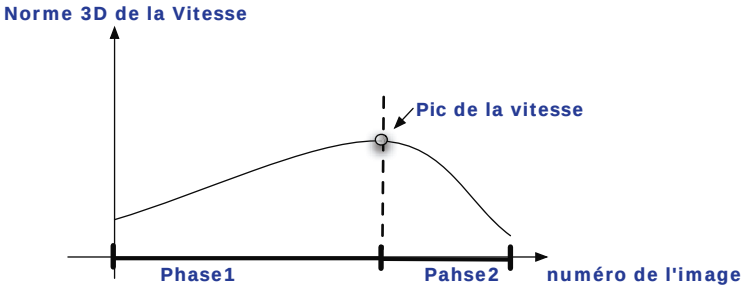


FIGURE *4.14 – Signature d'un pointé manuel Sign*1

– C2 : l'immobilité de la main qui réalise le signe et donc de son mouvement.

$\Rightarrow S2$ est le résultat de $C1 \wedge \neg C2$.

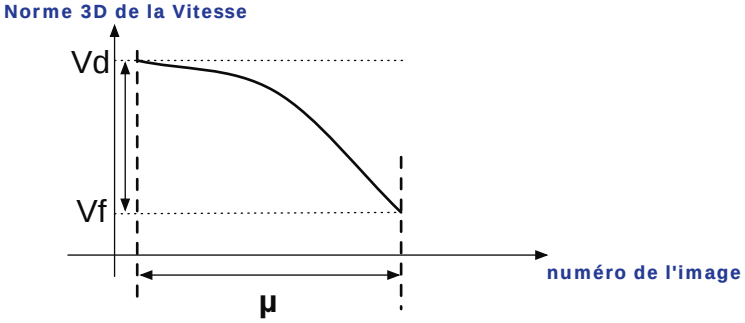


FIGURE *4.15 – Signature d'un pointé manuel Sign*2

A partir du corpus de capture de mouvement *« Marqspat »*, nous avons évalué les signatures des deux modèles de profil dynamique du signe de pointé manuel et obtenu les valeurs suivantes :

Pour $S1$:

– Une durée minimale de (2.24 - 0.73 ≈ deux images) pour la première phase.
– Une durée minimale de (1.51 - 0.94 ≈ une image) pour la deuxième phase.
– La durée totale ne dépasse pas 10 images.

Pour $S2$:

– Une durée minimale de (2.6 - 0.4 ≈ deux images) pour la pseudo-phase [7].
– La durée totale ne dépasse pas 10 images.

Nous avons évalué la validité des signatures $S1$ et $S2$ en construisant deux filtres basés sur la durée moyenne de chaque phase. Les filtres ont été appliqués sur des mesures de

7. Nous appelons pseudo-phase le début du geste marqué par un pseudo-palier

vitesses de la main droite effectuées sur le corpus *« DEGELS »* décrit dans le chapitre (3). Il s'agit d'une vidéo sur laquelle nous avons réalisé une annotation manuelle des positions de la main droite dans la vidéo en question. [8].

Parmi 16 pointés, nous avons enregistré les résultats de détections présentés dans le tableau (4.15) qui mettent en avant le taux faible de vraies détections et le nombre très élevé de faux positifs.

Signature	**VP**(%)	**FP**(%)
$S1$	56.25	$\approx$ 67.8
$S2$	12.5	$\approx$ 48.7

TABLE *4.15 – Résultats de détections de $S1$ et $S2$ dans une vidéo*

Le taux élevé de fausses détections montre que le profil dynamique est nécessaire mais pas suffisant pour caractériser un signe de pointé manuel qui représente un cas particulier du référencement. Ceci signifie que :

- La transformation des signatures de $3D$ en $2D$ par élimination de la composante de profondeur altère le modèle de départ.
- Il existe des profils des signes de pointé manuel non représentés par les signatures $S1$ et $S2$.

L'analyse des vitesses instantanées de la main droite lors de la réalisation de signe de pointé a révélé plusieurs profils. Nous avons choisi de représenter les profils les plus fréquents car on ne peut représenter tous les profils car le modèle résultant serait dépendant du corpus étudié. Nous avons mentionné dans le chapitre (2) qu'il existe plusieurs paramètres susceptibles de caractériser le geste de référencement. Ceci nous conduit au pas suivant qui consiste à introduire d'autres paramètres afin d'améliorer la robustesse du modèle de référencement.

A partir du profil de déplacement de la main (au cous d'un pointage) observé par [Dalle 2009], nous avons déduit la pertinence d'étudier le profil dynamique de pointage. Nous en retenons qu'il existe plusieurs profils dynamiques du mouvement de la main. Par conséquent, on ne peut représenter tous les profils. Nous avons gardé les profils les deux plus fréquents illustrés par les graphes 4.14 et 4.15.

4.4.4 Modèles

Nous nous sommes focalisées sur les aspects temporel, spatial et dynamique des gestes de référencement. Nous en retenons que :

8. L'annotation est manuelle et graphique à la fois car elle utilise une application de détection des clics de la souris

– pour l'aspect temporel, certaines relations entre motifs peuvent être représentées par une logique temporelle simple. Néanmoins, nous avons constaté qu'il existe d'autres types de relations plus complexes.
– pour l'aspect spatial, la combinaison des pointés (regard, tête et main) peut être considérée comme un nouvel indicateur de précision d'un même pointé vers un même locus.
– pour l'aspect dynamique, il existe plusieurs profils de vitesse.

Ces constatations motivent nos choix portés sur :

– Les classes les plus fréquentes de profils de vitesses de la main droite.
– La précision dans le signe de pointé interprétée à partir de l'évolution de la distances entre : 1) la main droite et un ou plusieurs locus, 2) la cible du regard et le locus, 3) la droite portant la direction de la tête et le locus.
– Les classes floues de relations temporelles entre gestes combinés.

Les règles qui découlent de ces propriétés seront utilisées dans le système de reconnaissance de référencement comme illustré dans le tableau suivant (4.16).

Aspect	**Propriété**	**Paramètre**
Temporel	Ordre chronologique des gestes	Décalage entre gestes
Spatial	Position de la MD, de la droite de la direction de la tête, de la cible du regard et du locus	Distances 3D entre composantes corporelles et locus
Vitesse	Profil dynamique de la MD	Durées de chaque phase

TABLE *4.16 – Les paramètres qui seront introduits dans le système de détection de référencement*

4.5 Conclusion

L'analyse des gestes de référencement que nous avons menée montre que la réalisation de référencement se caractérise par la multi-linéarité dans la majorité des cas rencontrés dans les corpus de modélisation *« SignCom »* et *« Marqspat »*. Nous avons montré dans le paragraphe précédent par l'exemple l'intérêt de l'aspect combinatoire dans la précision de l'événement de référencement en terme de distances relatives. Le choix de l'étude de tous les gestes qui contribuent au concept de référencement a été justifié, au départ, par une analyse préliminaire sur les profils dynamiques de la main droite qui a montré la nécessité d'introduire d'autres paramètres afin de rendre le modèle représentatif de

référencement robuste. Pour cela, nous avons séparé l'analyse de gestes de référencement en deux types : 1) temporelle : les délais entre gestes et 2) spatiale : les positions relatives des composantes corporelles par rapport aux loci.

Sur le plan temporel, le type de combinaisons gestuelles 'MNM' s'annonce conséquent en termes de durée et de nombre d'occurrences par rapport à la durée totale et le nombre de séquences de référencement dans un discours signé. De plus, nous avons remarqué une tendance vers le déploiement de combinaisons de degré 3.

Nous avons émis, dans le chapitre (2) une interrogation sur l'existence d'une relation entre le type de la zone référencée et le type de geste déployé pour le référencement manuel (pointé). Dans ce chapitre, nous avons utilisé les paramètre de proximité et de nombre de locus dans une zone référencée afin de donner des éléments de réponse. En effet, à partir du corpus *« Websourd »*, nous avons traité l'influence du nombre de loci occupant une zone sur la réalisation gestuelle de référencement et donc le comportement des composantes corporelles. Nous en retenons que la précision de la zone pointée ne dépend pas du nombre de loci occupant une zone. D'un autre côté, nous sommes persuadés que l'interlocuteur perçoit avec précision la zone à laquelle le locuteur avait l'intention de faire référence. Ainsi, nous nous interrogeons sur le ou les autres paramètres qui contribuent à une interprétation non ambigüe du locus référencé suite à un pointé manuel. Nous soulignons par ce fait que la précision dans la réalisation de gestes de référencement met en jeu les gestes et la disposition de l'espace de signation perçus par l'interlocuteur. Ceci nous conduit à suggérer une redéfinition de la notion de précision dans le référencement en fonction des gestes déployés et de la structure de l'espace à un instant donné.

Notre objectif est de mettre en place un système de reconnaissance du référencement. Etant donné que le système de reconnaissance prend en entrée des données bi-dimensionnelles, nous proposerons dans le chapitre suivant une méthode de passage $3D \Rightarrow 2D$ des modèles tri-dimensionnels décrits dans ce chapitre.

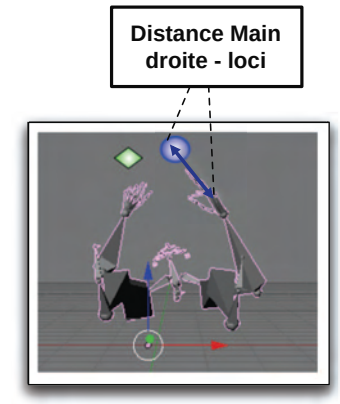
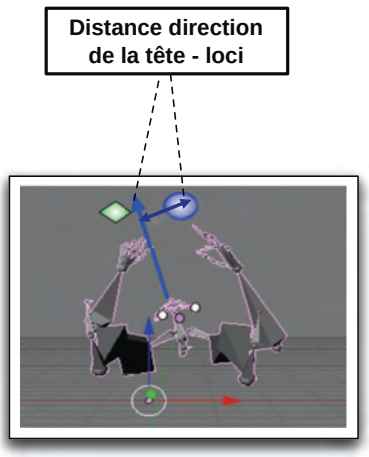
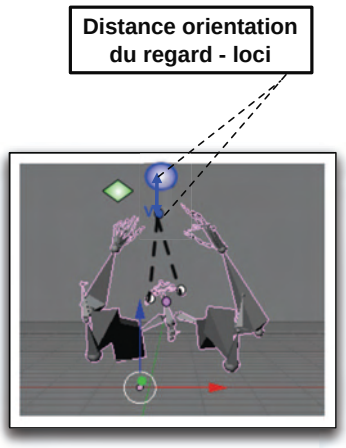
Signification	Modèle géométrique
	Distance tri-dimensionnelle séparant la sphère représentative de la main droite (4.2.2) du locus (4.2.5)
	Distance tri-dimensionnelle séparant la droite représentative de l'orientation de la tête (4.2.3) du locus (4.2.5)
	Distance tri-dimensionnelle séparant la cible du regard (4.2.4) du locus (4.2.5)

TABLE *4.13 – Vue de dessus de modèles géométriques*

CHAPITRE 5

Exploitation

Dans ce chapitre, nous décrirons l'intégration des résultats auxquels nous avons abouti dans un système de reconnaissance de structures de référencement. Nous proposons uneapplication d'apprentissage et de recnnaissance qui a pour but d'évaluer les modèles construits. Pour cela, nous passerons en revue les principales méthodes de reconnaissance existantes, puis proposerons le modèle de conception de la méthode de reconnaissance retenue. Prenant en compte les détails du corpus d'exploitation décrits dans le chapitre (3), nous conclurons par les résultats de reconnaissance de structure syntaxique de référencement dans une vidéo d'énoncé en langue des signes.

Sommaire

5.1 Problématique

On se propose d'intégrer, dans un système de reconnaissance automatique, des critères de nature temporelle et spatiale. On s'interroge, tout d'abord, sur les caractéristiques de la méthode de reconnaissance de structure de référencement qui aura pour entrée un flux vidéo d'énoncés en langue des signes et les modèles spatio-temporels de référencement. Le problème consiste à déterminer une méthode de reconnaissance qui prend en compte les critères correspondant :

- au profil de la vitesse $3D$ de la main droite,
- à la variation des distances $3D$ entre composantes corporelles et loci,
- au modèle temporel qui décrit l'enchaînement des gestes.

La méthode de reconnaissance fournit en sortie des segments de référencement auxquels on attribue des taux d'appartenance des mesures correspondantes aux classes de paramètres présentés et utilisés par les modèles (vitesse, distance et décalage temporel).

La mesure du taux d'appartenance implique des mesures 2*D* réalisées sur des images. Une homogénéisation des données 2D vidéo et des mesures 3D du modèle spatial est donc nécessaire. Nous étudions donc la méthode de transformation des données de type image et / ou sur le modèle spatial établi. Ceci portera sur le passage du modèle spatial tri-dimensionnel vers un modèle 2D. Rappelons que le modèle tri-dimensionnel est le résultat d'analyses de corpus dont les données sont nécessaires à la reconstruction complète de gestes de référencement sans biais (capture de mouvement et du regard, vidéos annotées). Le modèle 2D représente le résultat de transformation du modèle tri-dimensionnel. La fonction de transfert correspondante a pour objectif de fournir un modèle compréhensible par un programme de traitement de données qui lors de l'exploitation seront 2*D*.

Nous allons donc calculer un indice de similarité entre la classe représentative du modèle 2D 2*D* – résultat de transformation 2*D* du modèle 3*D* en question – et la classe des mesures réalisées sur des images 2*D* (*« SignCom »*). Nous nous interrogeons donc sur les classes que nous allons définir.

Ensuite, nous intègrerons cet indice de similarité dans le processus de classification. Dans la suite, nous passerons en revue les grandes familles de méthodes de classification statistiques. Nous détaillerons les méthodes appliquées aux données de natures spatiale et temporelle. Par la suite, nous présenterons brièvement les méthodes de transformation vers des modèles 2D. Le résultat d'intégration sera présenté sous la forme d'un diagramme de conception UML à partir duquel nous développerons un programme d'apprentissage et de reconnaissance.

5.2 Méthodes de classification

Le choix de la méthode de classification est vaste, pour cela, nous nous limiterons aux méthodes adaptées aux types de données dont nous disposons.

5.2.1 Choix préliminaire des méthodes

Les méthodes paramétriques requièrent la connaissance totale des densités de probabilité régissant la distribution des observations. Dans les applications réelles, ces fonctions sont représentées par des lois uniformes de Gauss, dont les paramètres sont estimés. L'estimation des paramètres se fait à partir des données d'apprentissage. Les méthodes paramétriques requièrent ainsi un volume conséquent de données d'apprentissage qui ne sont pas disponibles pour les corpus en langues des signes et en particulier les référencements dans un énoncé en langue des signes.

Le principe des systèmes experts est la modélisation de la connaissance à l'aide d'une base de connaissances sous la forme de règles. La base consiste en un ensemble de règles de deux types. Des règles bas niveau qui permettent de décider de l'état des ROIs (en mouvement ou pas) en se basant sur la mesure obtenue et d'interpréter la

nature du mouvement correspondant. Des règles de haut niveau qui consistent à associer ou pas un mouvement donné à une structure de référencement (Voir chapitre 2). Les règles bas niveau se basent uniquement sur le contenu de l'image. Les règles haut niveau sont générées à partir du modèle spatio-temporel et sont sensées aboutir à une décision binaire. Cependant, le modèle spatio-temporel se présente sous la forme de signatures gestuelles dont la certitude est pondérée (voir l'exemple 4.8). Il faudrait donc opter pour une méthode qui génère, dans un premier temps, des hypothèses pondérées et, en fonction des taux d'appartenance, génère une décision finale.

Les méthodes non paramétriques se basent sur l'estimation de fonctions d'appartenance. En particulier, les méthodes de classification floues permettent de représenter un ensemble de mesures (vitesses, distances, etc.) sous forme de classe associée à un modèle de mesures de même type. L'association entre classe et modèle est pondéré par un taux d'appartenance. Dans le domaine de la reconnaissance de signes dans un énoncé en langue des signes, certains travaux utilisent la fusion de classes de propriétés gestuelles de signes dans le but de reconnaître des signes manuels isolés. Par exemple, [Bedregal 2006] et [Phitakwinai 2008] utilisent la classification floue pour reconnaitre des gestes ou des lettres. [Hieu 2008] et [Futane 2012] ont utilisé des méthodes qui combinent à la fois la classification floue et d'autres techniques statistiques telles que les modèles de Markov cachés et les réseaux de neurones, dans le but de reconnaître des signes manuels isolés. [Al-Jarrah 2001] et [Holden 1999] ont mis en place des moteurs d'inférence et une base de règles de classification floue afin de décider de la signification des signes isolés. [Fang 2004] a mis en place un arbre de décision basé sur des classificateurs de caractéristiques gestuelles de signes réalisés par une seule main. [Sarfraz 2005] a utilisé le principe de la logique floue dans le but de développer un suivi des mains en mesurant des scores de similarité entre le modèle d'objet et la forme détectée.

Cet aperçu montre que les systèmes experts pourraient être adaptés à l'interprétation de résultats fournis par les opérateurs de mesure de propriétés gestuelles dans l'image (rotation, translation, mouvement rapide, lent, etc.). La classification floue serait une solution appropriée à la génération de classes de mesures représentant chacune une variante du modèle spatial et / ou temporel selon un taux d'appartenance donné. Nous construirons, ainsi, une méthode de reconnaissance qui se compose de deux étapes principales. La première concerne la traduction des mesures réalisées à partir de coordonnées $2D$ de zones d'intérêts segmentées dans une suite d'images d'un extrait de vidéo d'énoncé en langue des signes. Puis d'une étape de génération d'une liste d'intervalles temporels exprimant des variantes de référencement. Le système de reconnaissance global, que nous proposons, est illustré dans la figure(5.1).

Dans la suite de cette section, nous passerons en revue les détails des systèmes experts implémentés pour l'interprétation d'images, d'une manière générale, et pour l'interprétation ou pas des mesures $2D$ en des classes de mesures associées aux modèles de référencement.

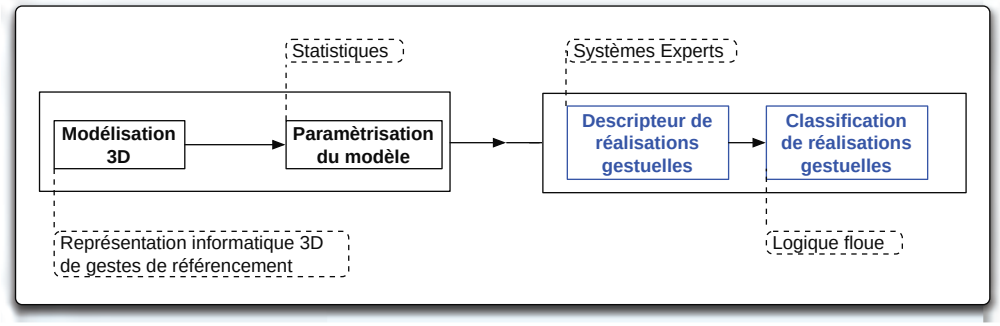


FIGURE *5.1 – Etapes de reconnaissance de structures de référencement*

5.2.2 Méthode de modélisation de la connaissance

D'une manière générale, l'analyse d'images par un système expert se réalise en trois étapes : 1) la génération du plan de l'analyse : cette tâche consiste à générer un guide pour l'analyse d'une image donnée en tenant compte des caractéristiques de l'image en question et la procédure classique d'analyse d'images : 2) la sélection d'opérateurs de traitement d'images et 3) l'ajustement de paramètres mis en jeu par l'analyse en question. Dans la littérature, on trouve des références sur l'automatisation de la segmentation d'images guidé par les systèmes experts, appliqué le plus souvent dans le domaine de l'imagerie médicale telles que la segmentation d'une image par raisonnance magnétique (IRM) [Pit 2004]. Nous nous sommes intéressé à la méthodologie utilisé pour mettre en lien la segmentation d'images et les systèmes experts. [Matsuyama 1988] décrit quatre systèmes de segmentation d'images par l'utilisation de système expert :

1. Système de consultation basé sur les spécifications fournies par l'utilisateur selon lesquelles il fournit une instanciation d'algorithmes de traitement d'images dont le résultat sera validé par l'utilisateur.
2. Système de composition de programmes basé sur des spécifications détaillées sous forme de figures ou de textes qui, une fois traduites en une suite d'instructions d'opérateurs et appliquées sur l'image d'origine, seront validées par l'utilisateur.
3. Système de conception d'algorithme de segmentation basé sur des règles de décision dont le rôle est de guider l'usage des opérateurs de traitement d'images vers un résultat optimal.
4. Système de segmentation guidé par l'objectif moyennant des processus de transformation qui facilitent la sélection d'opérateurs appropriés. L'objectif formulé par l'utilisateur sert à spécifier les opérateurs bas niveau et à vérifier la cohérence du résultat fourni par le système grâce à deux analyses parallèles *« up-donw »* et *« bottom-up »*.

Ce travail discute les méthodes de génération de plan d'exécution et de la suite des opérateurs à appliquer ainsi que les méthodes de validation des résultats. Il en ressort que les systèmes (1), (2) et (4) sont basés sur des primitives d'images ce qui les rend moins flexibles et surtout ne gèrent pas des données externes comme le contexte et l'historique du discours, ainsi que le modèle spatial. Par contre la méthode de validation de résultats dans le système (4) propose un processus de vérification plus flexible car elle permet la prise en compte de l'historique, du contexte et du modèle spatio-temporel. Dans le système (3), la méthode de génération de plan d'exécution et d'instanciation de paramètres sont génériques et peuvent gérer le contenu temporel dans une suite d'images comme le requièrent les spécification de la caractéristique temporelle du modèle de référencement.

Ce que l'on propose c'est un système de décision bas niveau qui gère l'historique d'images ainsi que les données interprétées provenant de fonctions de haut niveau. Le pseudo-algorithme (1) est un exemple de reformulation de règles de décision bas niveau concernant le contenu d'une image donnée dans le but de déterminer les régions d'intérêt (ROI) détectées.

Algorithm 1 Règles conditionnelles *SI .. ALORS*

1 : **if** $PixelVal \in [val_1..val_2]$ & $PixelCoordinates \in [(x_1, y_1)..(x_2, y_2)]$ **then**
2 : $\quad MERGE(Roi, Pixel)$
3 : **end if**
4 : **if** $RoiSize \in [size_1..size_2]$ **then**
5 : $\quad FOCUS(Roi)$
6 : $\quad MEASURE\_ROTATION(Roi)$
7 : **end if**

Comme l'illustre cet exemple, les règles de décisions sont appliquées à deux niveaux de type de données, pixel et région d'intérêt. L'algorithme de décision prend en entrée :

- L'historique du discours.
- L'historique des réalisations gestuelles.
- Les modèles spatio-temporels de référencement.

Et fournit en sortie l'appartenance d'un pixel donné (à une région d'intérêt).

Les décisions sous forme d'interprétations se basent sur l'historique du discours pour décider d'une hypothèse d'un référencement. Rappelons que le référencement est un concept linguistique lié aux fonctions linguistiques énoncées dans le chapitre (2) : Le signe de pointé, le signe locatif, le signe en mouvement et le signe localisé.

L'historique des réalisations gestuelles représente une donnée d'apprentissage au système de reconnaissance. En effet, le modèle $2D$ fournit des décisions bas niveau sur la similarité entre mesures $2D$ et modèle $2D$ de référencement.

Nous verrons dans la section suivante les spécifications de la méthode de classification des mesures $2D$.

5.2.3 Méthode de classification floue

Comme nous l'avons vu dans la sous-section (5.2.1), l'utilisation de la classification floue a pour objectif la reconnaissance de signes manuels isolés. Notre travail s'étend sur le traitement de vidéos de discours continu en langue des signes. Aussi, nous nous intéressons aux différents canaux gestuels (mains, tête, regard, etc.) et donc aux traitements ordonnancés selon les modèles temporels. Nous nous intéressons également à la localisation des zones spécifiques de l'image (main droite et tête). Nous avons établi un bilan des travaux de classifications floues qui tiennent compte de données hétérogènes multi-sources de natures spatiales et temporelles. [Guillemot 1999] a développé une méthode de classification qui gère les classes multi-modales et tient compte des possibilités de superpositions de classes qui fournissent, pour chaque observation, un degré d'appartenance à chaque classe. [Jadon 2001] a développé un système de détection d'interruptions d'activités dans une vidéo moyennant une interprétation de haut niveau de changements inter-images. En langues des signes, [Holden 1996] a développé une méthode de reconnaissance de signes isolés basée sur la description classique d'un signe (emplacement, configuration, orientation et mouvement). Un degré d'appartenance est estimé à partir des configurations clés puis vient la décision du signe en fonction du mouvement réalisé entre les deux postures clé. [Sarfraz 2005] a développé une méthode de calcul de similarités de deux surfaces des deux mains réalisant un signe réalisé par les deux mains. [Bedregal 2006] a développé une méthode de classification de configurations manuelles à partir de données de capture de mouvement moyennant l'estimation de degré d'appartenance de configurations à des classes représentatives des configurations manuelles.

Sur le plan temporel, [Badaloni 2000] a proposé une représentation floue des relations temporelles d'Allen (Voir tableau 4.2). En effet, il a introduit un coefficient d'appartenance à chaque type de relation temporelle formalisé comme suit :

$$I_1(Rel_1[\alpha_1], Rel_2[\alpha_2], ..., Rel_{13}[\alpha_{13}])I_2 \tag{5.1}$$

Tels que :

I_1 et I_2 représentent des intervalles temporels.

$Rel_i(\alpha i)$; $(i = 1, 2.., 13)$

α_i est le coefficient d'appartenance ou degré de préférence appartenant à l'intervalle [0..1].

[Badaloni 2000] a également introduit les opérateurs de conjonction et de disjonction dans le but de réduire le nombre de relations temporelles possibles (5.1). Il s'agit de la méthode « IA^{fuz} ».

Soient deux relations $R^{'}$ et $R^{"}$, la relation temporelle résultante est définie comme suit :

Opérateur de conjonction :

$$R = R^{'} \otimes R^{"} = (Rel_1[\alpha_1], Rel_2[\alpha_2], ..., Rel_{13}[\alpha_{13}]) \quad (5.2)$$
$$\alpha_i = min\{\alpha_i^{'}, \alpha_i^{"}\}; i \in \{1, .., 13\}$$

Opérateur de disjonction :

$$R = R^{'} \oplus R^{"} = (Rel_1[\alpha_1], Rel_2[\alpha_2], ..., Rel_{13}[\alpha_{13}]) \quad (5.3)$$
$$\alpha_i = max\{\alpha_i^{'}, \alpha_i^{"}\}; i \in \{1, .., 13\}$$

5.2.4 Synthèse

L'intégration du modèle tri-dimensionnel de référencement figure parmi les étapes de la méthode de reconnaissance proposée (Voir figure 5.1).

Nous avons mentionné que l'implémentation du modèle statistique présenté dans le chapitre précédent nécessite un passage vers un modèle 2D qui relate la nature des données d'entrée c'est-à dire la vidéo. Le passage d'une représentation $3D$ à une représentation $2D$ s'inscrit dans le cadre de transformation de modèles. Nous avons constaté que certains travaux dans le domaine de l'ingénierie des logiciels ont abordé le thème général de transformation de modèles appelés modèles d'exigences. Nous nous sommes intéressés à la conclusion déduite du travail de [Kaindl 2006] qui mentionne qu'il est possible d'effectuer une transformation de modèles moyennant une mise en correspondance (appariement) des spécifications des deux modèles. Cette théorie nous a permis de vérifier que le passage $3D \rightarrow 2D$ est réalisable.

D'autre part, il existe plusieurs travaux opérant dans le cadre de reconstruction de données tri-dimensionnelles à partir de données bi-dimensionnelles dont nous citons ceux qui se focalisent sur l'estimation de la posture d'une personne. [Lefebvre-Albaret 2010] présente une amélioration des méthodes de reconstruction de postures adaptées à une personne qui signe. Il a établi un ensemble d'hypothèses sur les mouvements d'un signeur et des conditions d'enregistrement dans le but d'optimiser la reconstruction de la posture. Il a ensuite passé en revue les méthodes logicielles qui permettent la reconstruction de la posture.

Nous avons choisi de conserver ces conditions d'enregistrement en raison des résultats prometteurs auxquels le travail a abouti. Les conditions et hypothèses sont :

- Sujet restant seul dans l'espace de travail et constamment filmé de face.
- Vêtements colorés spéciaux et proches du corps.
- Pose de départ connue.
- Caméra statique ou avec un déplacement connu.
- Fond fixe et uniforme.

Nous nous permettons d'ajouter à ces conditions l'hypothèse que le mouvement de la tête est caractérisé, dans la plupart des énoncés en langue des signes, par la lenteur.

En ce qui concerne les méthodes de reconstruction d'une figure géométrique à partir d'une vidéo, [Lefebvre-Albaret 2010] en a présenté les principales catégories de méthodes génériques en fonction de l'information exploitée et le résultat fourni. [Lefebvre-Albaret 2010] a pointé les inconvénients de chaque catégorie par rapport à la nature des données vidéo en langues en langue des signes. Ensuite, il a proposé une nouvelle approche adaptée à ce type de données qui tient compte des hypothèses et conditions établies précédemment. [Tomasi 1992] a utilisé la méthode de *« Optic flow »* qui fait partie de la catégorie des outils *« shape from motion »*. Cette méthode a l'avantage de récupérer la trace de mouvement d'un objet même si ce dernier disparaît pendant un moment de l'espace de travail capturé. Cependant, la méthode *« Optical flow »* est la reconstruction $3D$ à partir de plusieurs vues d'une capture. Nous n'avons pas pu mettre en oeuvre cette méthode car nous ne disposons pas des vues nécessaires dans les différents corpus. Dans la section suivante, nous mettrons le lien entre la méthodologie de reconnaissance adoptée et le modèle de référencement construit.

5.3 Mise en oeuvre

Nous avons choisi une méthodologie de reconnaissance qui prend en compte le modèle 3D de référencement et les corpus composés de vidéos. Dans cette section, nous affinerons la description de l'approche de reconnaissance adoptée, les étapes de construction de connaissances et les modules correspondants.

5.3.1 Approche adoptée

Dans ce paragraphe, nous listons les méthodes utilisées par le système de reconnaissance que nous avons mis en place. L'approche adoptée dans la construction d'un système de reconnaissance. Nous avons choisi deux approches que nous avons combiné par la suite. Une approche ascendante qui prend en entrée un ensemble d'images et fournit en sortie des mesures de distances et de vitesses des zones d'intérêt et une approche descendante qui prend en entrée le modèle 3D de référencement et fournit en sortie un coefficient de détection du référencement dans cet ensemble d'images. Dans la suite, nous présenterons le processus de la méthode de reconnaissance sous forme de schéma et détaillerons le rôle de chaque étape ainsi que les données traitées.

5.3.1.1 Approche ascendante

Segmentation ROI $< Si...Alors >$ sont utilisés pour décider si un ensemble de pixels représente une ROI et pour décider de la nature de la ROI.

Interprétation de mouvement $3D$ est utilisée pour déterminer le type de mouvement $3D$ à partir de la donnée de profondeur estimée.

5.3.1.2 Approche descendante

Le *« mapping »* est réalisé pour mettre en relation les propriétés du modèle spatio-temporel $3D$ et modèle $2D$.

Segmentation ROI sont utilisés pour décider si le mouvement estimé est similaire à celui décrit par le modèle gestuel d'entrée.

Mesure de coefficient de détection est utilisée pour attribuer une mesure de confiance à la similarité déduite selon le score de similarité et le résultat de vérification.

5.3.2 Types de données

5.3.2.1 Capture du regard

Nous avons mentionné dans le chapitre précédent que les données de capture du regard se présentent sous deux formes, enregistrées et temps réel. Nous proposons de prendre en compte les deux types de données dans le processus de reconnaissance car elles sont de même type – les coordonnées tri-dimensionnelles de la cible du regard, seule la méthode d'interrogation de données change.

Capture en temps réel : Ces données sont fournies directement par l'équipement *« Facelab »* que nous avons présenté dans le chapitre (3). L'envoi de la requête d'interrogation de l'angle d'orientation du regard instantané et la réponse correspondante sont réalisées grâce à la librairie *« Boost »* installée dans un poste client qui forme avec l'équipement *« Facelab »* un réseau local.

Capture enregistrée : L'interrogation de données enregistrées consiste en une lecture automatique des ressources à accès direct (sous forme de tableurs).

5.3.2.2 Vidéos

De la même manière, les données vidéo se présentent sous deux formes, enregistrées ou flux capturé en temps réel par une caméra. Nous avons traité seulement les vidéos enregistrées car nous ne gérons pas la sauvegarde automatique d'images pour pouvoir les traiter en temps réel. La structure d'image est désignée par un pointeur sur un type *image*. Nous avons choisi la librairie de traitement d'images *« OpenCV »* qui intègre ce type de structures.

5.3.3 Bases de connaissances

La base de connaissance se compose de deux types d'information :

Connaissances de type spatial qui représente le résultat de conversion du modèle tri-dimensionnel de la variation de la distance entre ROIs.

Connaissances de type temporel qui représente le modèle de décalage temporel formalisé dans le chapitre précédent.

Connaissances de type dynamique qui représente le résultat de conversion du modèle tri-dimensionnel du profil de vitesse de la main droite.

5.3.4 Modules

5.3.4.1 Traitement de modèles

Les fonctions listées ci-dessous permettent la traduction du modèle spatial $3D$ de référencement sous forme de règles. Le résultat représentera donc la base de connaissances.

Projection de modèles (1) : Il s'agit de la transformation des modèles tri-dimensionnels établis dans le chapitre (4) en un modèle $2D$. Cette solution admet implicitement que :

- La capture de l'énoncé est réalisée par une seule caméra statique ce qui est le cas.
- Le signeur ne change pas de position au cours de l'énoncé

La transformation de modèles consiste en la suppression de la donnée de profondeur ce qui donnera un modèle $2D$. Le modèle obtenu consiste en : 1) la variation de distances entre la région main droite ou la tête et la zone locus qui est représentée également par la zone main à un instant antérieur (localisation de signe). Cela revient à mesurer la distance $2D$ entre la zone main droite et la zone tête, et 2) la variation de l'amplitude de la vitesse $2D$ de la main droite.

Mise à jour du modèle $2D$ **(2) :** Il s'agit de l'ajout de mesures associées à un paramètre du modèle spatial bi-dimensionnel (distance et / ou vitesse). L'ajustement du modèle $2D$ se fait manuellement quand il s'agit de faux négatifs (référencements non détectés).

5.3.4.2 Traitement de vidéos

Similarité temporelle (3) : Cette étape consiste à estimer les relations temporelles entre mouvements de la main droite, de la tête et du regard et en déduire la relation réduite selon la logique d'Allen comme cela a été expliquée dans le chapitre (4). La similarité représente le taux d'appartenance de la relation déduite fournie par le modèle temporel d'entrée et est notée « k_temp ».

Filtrage de mouvements (5) : En se basant sur l'hypothèse de lenteur du mouvement de la tête énoncée à la section (5.2.4), nous éliminerons les déplacements de la tête dont la variation du rapport : $\frac{Deplacement}{temps}$ est au-dessus d'un certain seuil. Le déplacement se traduit par une rotation en $3D$ et par un déplacement en $2D$.

5.3.4.3 Similarité spatiale $2D$ (7) :

Cette tâche consiste à mesurer la similarité entre des interprétations réalisées sur une séquence d'images et le modèle bi-dimensionnel résultant de l'étape (5.3.4.1). L'indice de similarité est noté k_spat et concerne les propriétés suivantes :

- La signature d'un mouvement manuel de signe de pointé.
- Les distances $2D$: 1) ROI main droite à un instant t – ROI main droite à un instant $t-1$ représentant le locus ou loci. 2) ROI tête – Main droite.

5.3.4.4 Décisions

Détermination de régions d'intérêts (4) : Il s'agit d'une étape qui permet de déterminer les pixels qui font partie des ROI de la tête et de la main droite. On se base sur le seuillage des valeurs et des positions des pixels pour déterminer a) les pixels de peau donc les

ROIs et b) la position des ROIs dont les zones tête et main droite. Cette opération se réalise une seule fois pour déterminer la position de la tête. Par contre, on fera appel à cette fonction pour estimer la position de la main droite sur une séquence d'images consécutives.

Estimation du taux de confiance (8) : Il s'agit de calculer le produit des deux indices de similarité temporelle et spatiale (voir diagramme 5.2).

Décision de référencement (9) : La fonction de décision fournit une réponse binaire (Référencement ou pas) selon les taux de confiance résultats des étapes (5.3.4.4) et (5.3.4.3).

5.4 Implémentations et résultats

En se basant sur le diagramme (5.2), nous proposons le diagramme de conception (5.3).

Il s'agit d'une vue conceptuelle simplifiée. Nous en avons exclu le module de synchronisation que nous avons évoqué dans le chapitre précédent. A partir du diagramme de conception UMLl (5.3), nous avons réalisé un module dans un langage orienté objet *« C++ »* compatible avec la bibliothèque *« OpenCV »* et une interface graphique *« Cocoa »* incluse dans l'environnement de développement *« Xcode »*. Nous avons utilisé un corpus vidéo enregistré dans les conditions citées dans la section (5.2.4).

Dans la suite, nous présenterons le pré-traitement des données vidéo, puis présenterons les résultats obtenus par le programme de reconnaissance appliqué au corpus d'exploitation décrit dans le chapitre (3).

5.4.1 Pré-traitement

Le pré-traitement permet d'évaluer la segmentation des ROIs et déterminer le biais entre zone projetée sur l'écran et zone pointée par la main droite, le regard et / ou par la tête. L'objectif est donc de trouver les ROIs main droite et tête et d'établir une relation entre la zone pointée sur l'écran (par la tête ou par la main droite) et la position réelle de la tête et de la main au même moment. De plus, nous nous proposons d'affiner le modèle $2D$ – qui représente le modèle $3D$ duquel on a enlevé la composante de profondeur – en réalisant les mesures de distances entre ROIs et de vitesses de la ROI main droite dans le cas de référencement. L'objectif du pré-traitement requiert donc un nombre conséquent de référencements ce qui n'est pas le cas pour les corpus de modélisation. Pour cela, nous avons mis en place une phase de calibrage qui précède l'énoncé en langue des signes (Voir description dans le chapitre 3).

5.4.1.1 Corpus de pré-traitement

La phase de calibrage se compose de deux étapes :

1. Pointer par le regard les mires qui apparaissent sur les coins de l'écran.

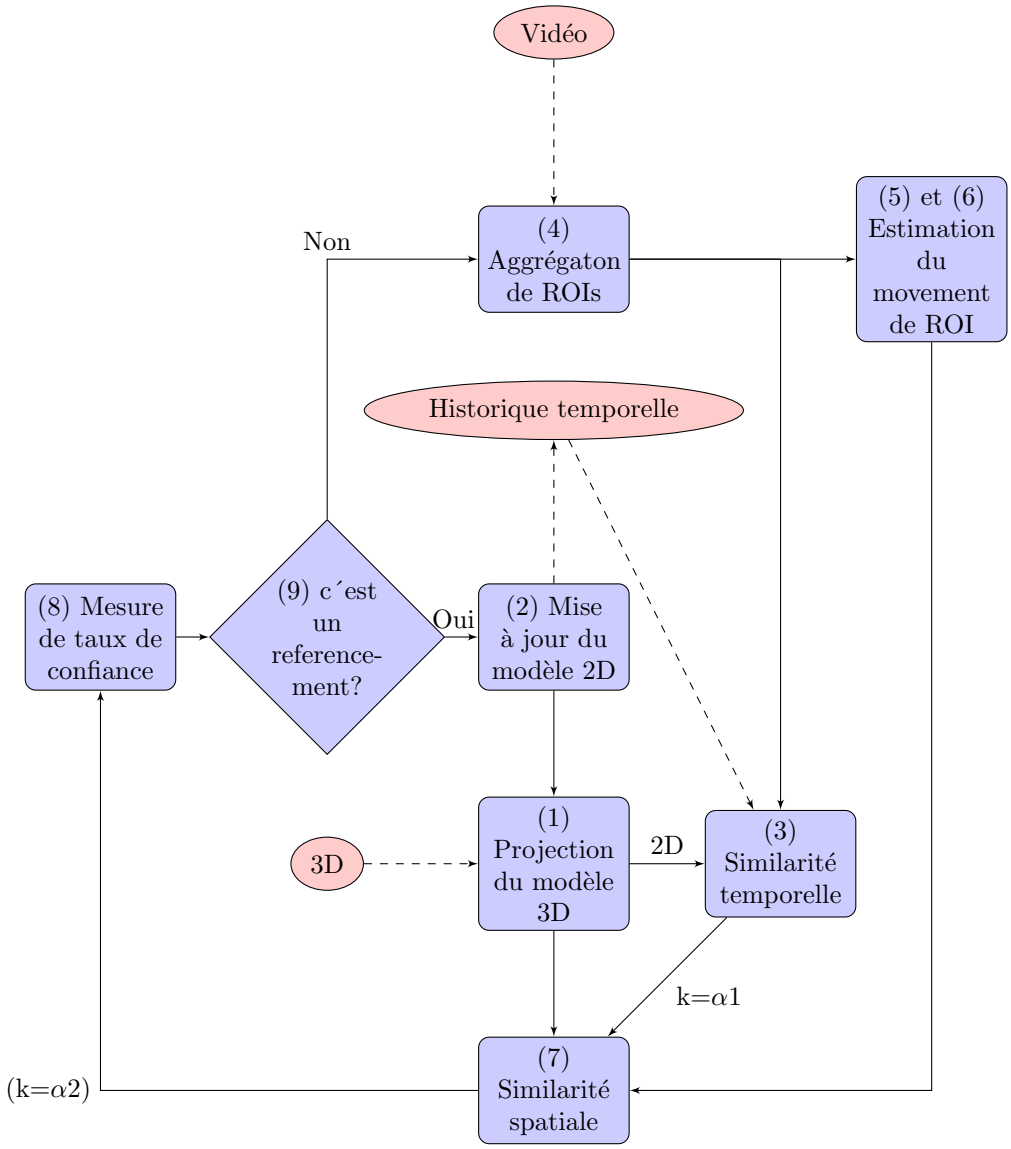


FIGURE *5.2 – Diagramme du processus de reconnaissance*

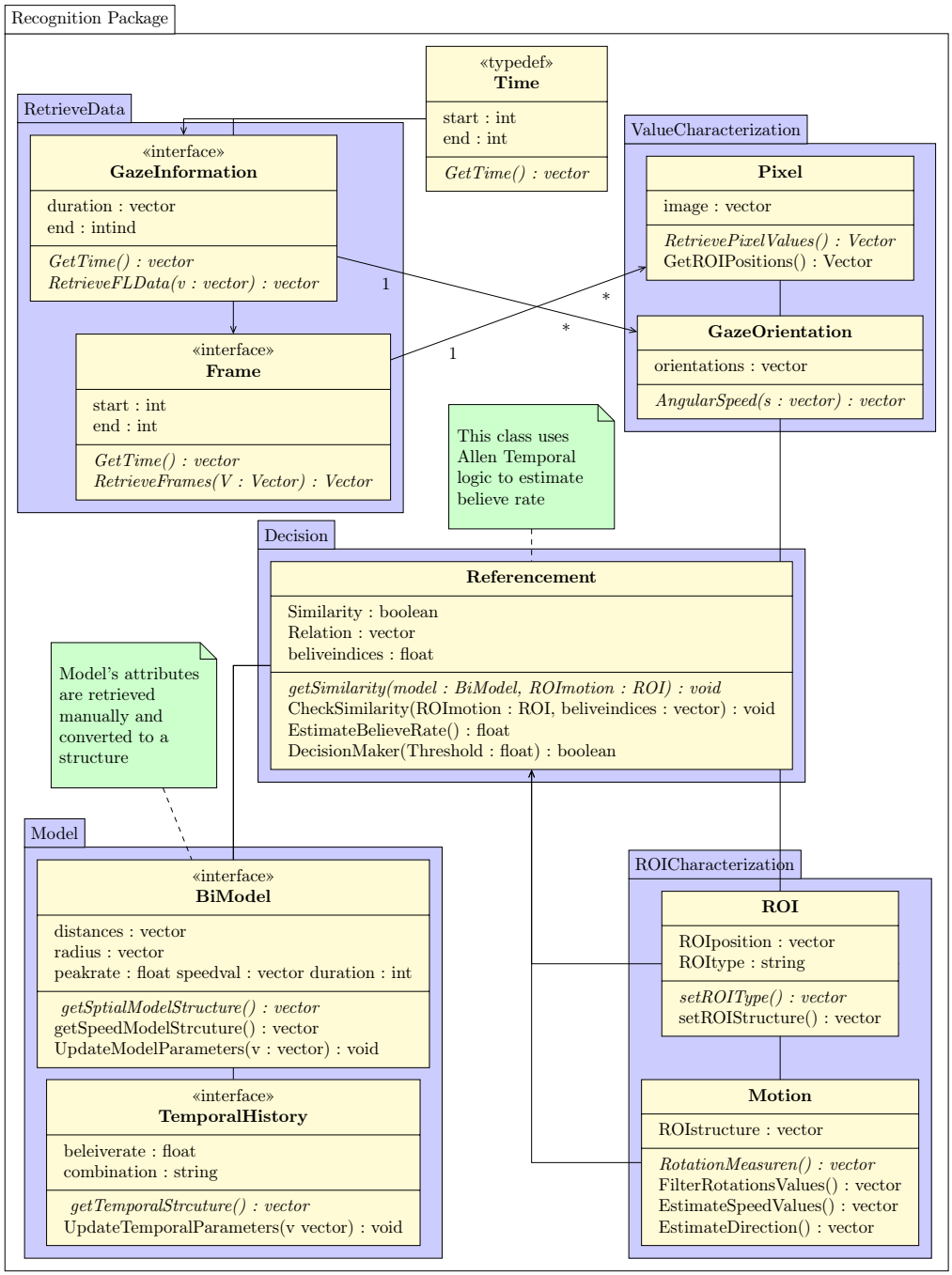


FIGURE 5.3 – *Diagramme UML du module de reconnaissance de référencement*

2. Pointer par la main des rectangles qui apparaissent aléatoirement sur l'écran.

Cette phase a été réalisée au cours de la captation du corpus décrit dans le chapitre (3) en langue des signes française (LSF), dans le cadre du projet ANR *« SignCom »*.

5.4.1.2 Segmentation des ROIs

Comme nous l'avons mentionné, le module de calibrage permet de segmenter les ROIs en filtrant les valeurs et positions des pixels. Les seuils sont fixés manuellement au début du programme. La valeur du pixel sélectionné appartient à un intervalle de couleurs de peau. La position du pixel sélectionné appartient à une zone de taille et de position connues. La zone est le résultat de découpage de l'image en zones Tête et Mains de la manière suivante :

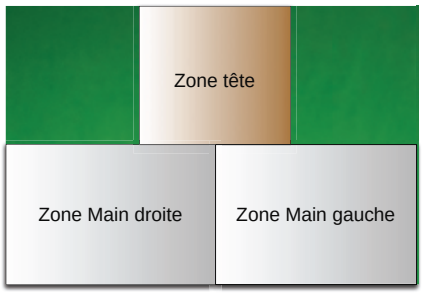


FIGURE *5.4 – Découpage de l'image en zones d'intérêt*

Les images (5.5) montrent le résultat de seuillage des valeurs de pixels des régions d'intérêt qui représentent la tête et la main droite.

FIGURE *5.5 – Résultat de seuillage des valeurs de pixels. Les zones rouges représentent les zones de pixels peau correspondants à un intervalle fixé (I). Les zones jaunes représentent les pixels de couleur plus claire que l'intervalle (I).*

Les images (5.6) montrent le résultat de segmentation des régions d'intérêt qui représentent la tête et la main droite dans une séquence de signes de pointés manuels.

FIGURE 5.6 – *Résultat de segmentation et de cadrage des ROIs selon les valeurs de pixels*

5.4.1.3 Les mesures $2D$

Cette partie est consacrée à la quantification de la variation de :

- la distance 2D entre ROIs,
- la vitesse de la ROI main droite.

La distance entre ROIs : Cette mesure représente le résultat de passage $3D \rightarrow 2D$ de la distance locus – tête et locus – main droite. La figure (5.7) représente l'évolution de la distance entre ROIs lors d'une séquence de signes de pointés manuels. Nous y observons des amplitudes variées (au nombre de 11) qui correspondent aux pointés manuels de la séquence. Nous avons choisi de représenter ces classes par des distances normalisées : $R = \frac{D_{T;M}}{D_{Tc;Mc}}$ tels que :

$D_{T;M}$ est la distance entre ROIs mesurée à chaque instant.

$D_{Tc;Mc}$ est la distance entre ROIs en phase de repos (signeur immobile).

On peut caractériser ces amplitudes par deux classes :

1. $C1$: Les pointés distants tel que $R = 1.09$
2. $C2$: Les pointés proches tel que $R = 0.86$

D'autre part, le contact entre les deux régions d'intérêt (la tête et la main droite) peuvent se traduire par l'inégalité suivante :

$$D_{ROI} <= d_M + d_T \tag{5.4}$$

D_{ROI} représente la distance entre les centres des deux ROIs.

d_M représente une estimation du rayon de la ROI main droite.

d_m représente une estimation du rayon de la ROI tête.

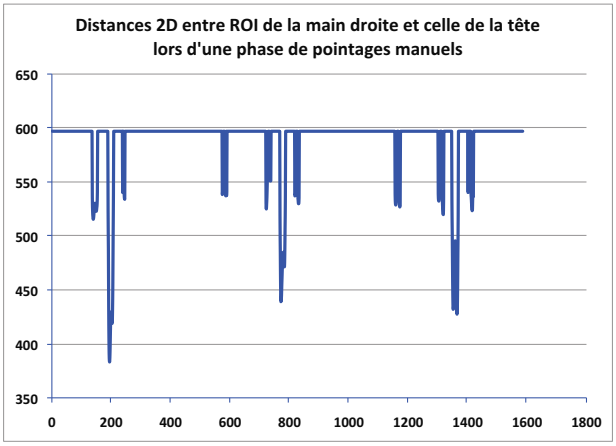


FIGURE *5.7 – Evolution de la distance entre ROIs lors de toute la séquence de calibrage*

L'inégalité (5.4) n'a pas été observée dans les mesures de distances entre ROIs effectuées sur les vidéos du corpus d'exploitation. Ceci signifie qu'aucune occultation n'a été observée dans les séquences de vidéos du corpus d'exploitation.

5.4.1.4 Choix de méthodes

Cette partie présente les solutions apportées aux problèmes detransformation de modèles 3*D* en 2*D*. Nous nous sommes interrogés sur les méthodes appropriées pour réaliser : a) la projection de modèles 3*D* et b) la segmentation temporelle de référencement par le regard utilisant les mouvements de la tête.

Quel(s) type(s) de projection ? Afin de rendre compréhensible le modèle 3*D* de variation de distance entre composantes corporelles et loci par rapport au système de reconnaissance, nous avons proposé dans (5.3.1.2) de procéder à un *« mapping »* entre les paramètres 3*D* avec leur équivalent en 2*D* comme suit :

- distance 3*D* → Distance 2*D*,
- sphère représentative de la main droite ou du locus → la zone *Main droite* 2*D* obtenue après segmentation des ROIs 2*D*,
- droite portant l'orientation de la tête → la position 2*D* de la ROI *Tête*.

En ce qui concerne le modèle de comportement, nous avons choisi de prendre en considération les variations de la distance bi-dimensionnelle qui sépare les zones d'intérêt (la main droite et la tête). Par ceci, nous émettons l'hypothèse que nous pouvons conserver les propriétés du modèle 2D telles que nous les avons spécifiées dans le chapitre précédent en ne tenant compte que des positions relatives.

Quelle méthode de segmentation temporelle ? Les mouvements de la tête sont estimés à partir de déplacement de la ROI en question. Les extrémas indiquent le moment de la stabilisation de la direction du regard. Nous admettons l'hypothèse que le pic d'amplitude

correspond à une fixation du regard. Ainsi nous considérons que le calcul d'extrémum local peut être une solution pour repérer les fixations et donc un indice de référencement par le regard.

Le fait que, l'enchaînement de projection de zones à pointer (dans les quatre coins de l'écran) impose à la tête des rotations consécutives de sens connus à l'avance, facilite le repérage des moments de repos de la tête et donc la déduction des instants correspondants aux fixations par le regard.

Le calcul des positions d'extrémums locaux dans l'évolution du comportement spatial de la main droite et de la tête est une solution adaptée au repérage de moments clés dans le profil de vitesses instantanées.

5.4.2 Test

Les solutions apportées pour la projection ($3D \rightarrow 2D$) et la segmentation temporelle mettent en jeu les paramètres résumés dans le tableau (5.1).

Procédure	**Caractéristique**
Projection $3D \rightarrow 2D$	Pointés manuels distants
	Pointés manuels proches
Segmentation temporelle	Angle de rotation de la tête

TABLE *5.1 – Caractéristiques apportées par les procédures de projection et de calibrage*

Nous nous proposons de prendre en compte ces paramètres en établissant le bilan de détection de Pointés (manuel et non manuel) dans un énoncé en langue des signes sous forme de flux vidéo. Nous avons exclu le traitement des vidéos des exercices sur la thématique de recettes car le nombre de Pointés et de production gestuelle, d'une manière générale est très faible. Toutefois, nous avons utilisé les sessions d'enregistrement de descriptions d'images. Il s'agit de 3 vidéos de descriptions de deux images (Voir illustration 3.10) d'une durée totale de 1 minute et 4 secondes. Les résultats des test se présentent sous forme de segments temporels d'une vidéo appartenant à un référencement. Nous avons obtenu les taux de vrais positifs (VP) résumés dans le tableau (5.2).

Vidéo	VP(%)	Classe	Types de relations	Taux correspondant(%)
Vidéo2	81.25	$C2$	$=$	50
			s	18.75
			f	6.25
			d	6.25
Vidéo3	80	$C2$	$=$	20
			s	60

TABLE *5.2 – Taux de détection de (VP) et étiquetage des décalages existants entre (VP) et référencement existant dans le cas de classement* $C2$

Nous avons joint aux résultats de détection, les relations temporelles qui lient les segments (VP) à la vérité terrain (VT). Nous rappelons les relations temporelles applicables dans ce cas dans le tableau (5.3).

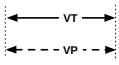
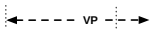
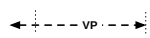
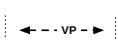
Etiquette	Illustration graphique	Représentation graphique
$'='$		$N_{VP} - (=) \rightarrow N_{VT}$
$'s'$		$N_{VP} - (s) \rightarrow N_{VT}$
$'f'$		$N_{VP} - (f) \rightarrow N_{VT}$
$'d'$		$N_{VP} - (d) \rightarrow N_{VT}$

TABLE *5.3 – Les relations temporelles représentatives des décalages temporels entre intervalles de référencement réels (VT) et l'intervalle de référencement détecté (VP)*

5.5 Conclusion

Nous sommes partis des données bas niveau (pixel) et haut niveau (modèle tridimensionnel de référencement) du corpus. Afin d'homogénéiser les deux types de données, nous avons opté pour une projection du modèle tri-dimensionnel par élimin ation de la composante de profondeur. La projection du modèle statistique nous a permis d'utiliser les paramètres du modèle $3D$: la distance entre locus et composantes corporelles ainsi que la vitesse de la main droite.

Ce que nous en retenons c'est que la tête reste immobile lors de la phase de pointage par le regard (d'après le modèle 2D). Cette constatation ne vérifie pas le modèle linguistique du pointé (un mouvement de la tête dans certains cas) qui a été établi à partir de la

vérité terrain. Ceci nous mène à déduire que la perception d'un mouvement de la tête ne correspond pas à un mouvement de la ROI tête (telle que nous l'avons défini) mais plutôt d'autres indices.

Nous avons testé, séparément, les stratégies de reconnaissance exposées dans ce chapitre et constaté l'existence de faux positifs (FP). Nous nous proposons dans le chapitre suivant d'expliquer le nombre important de fausses détections en faisant le bilan des limites de la technique de reconnaissance choisie ainsi que celles des corpus et modèles utilisés. Nous conclurons par une évaluation globale de la méthodologie de reconnaissance mise en place.

CHAPITRE 6

Evaluations

Dans ce chapitre, nous présentons une évaluation de la méthodologie de reconnaissance retenue ainsi que les raisons pour lesquelles nous avons obtenu des faux positifs dans les résultats de détection de segments de référencements. Nous évaluerons les entrées et les sorties de l'algorithme ainsi que les étapes de reconnaissance dans le but de : 1) pointer les limites du modèle 2D établi dans le chapitre (5) comme étant une des deux entrées de l'algorithme, 2) Remettre en question le modèle spatio-temporel construit dans le chapitre (4) et exploité dans le chapitre (5).

Sommaire

6.1 Introduction

La stratégie de reconnaissance retenue prend en entrée plusieurs modèles :
- des modèles de profils de vitesse de la tête et de la main droite,
- un modèle de décalage temporel entre gestes de référencement,
- un modèle de la variations des distances qui séparent la tête et la main droite du locus.

L'algorithme de reconnaissance fournit en sortie des intervalles temporels correspondants à des séquences vidéo dont les mesures (vitesses et distances $2D$) appartiennent à des classes qui correspondent aux critères décrits par les modèles cités ci-dessus. Les mesures sont classées selon le type (vitesse $2D$, distance $2D$ ou durée) et selon leurs taux de similarité aux modèles 2D. A partir de cette classification, nous établissons une décision finale sur la nature de la séquence (référencement / pas référencement). Nous nous proposons, dans ce chapitre, d'analyser les résultats de la classification obtenue et

d'évaluer les étapes de l'algorithme de reconnaissance retenu. Pour cela, nous avons opté pour la méthode d'évaluation *« VEST »* qui consiste à valider les entrées, les sorties et les étapes de l'algorithme. Cette méthode est appelée en anglais *« ETVX : Entry Task Validation Exit »* et a été définie par [Radice 1985] pour identifier et valider les critères d'un programme, ses tâches, etc. Nous utiliserons cette méthode afin de :

- Vérifier les niveaux de difficulté des données d'entrée par rapport à la capacité de l'algorithme mis en place à gérer les problèmes de traitement de vidéos en langue des signes.
- Vérifier l'exactitude des données de sortie en les comparant avec la vérité terrain.
- Evaluer les étapes de l'algorithme de reconnaissance, en particulier, détailler l'utilisation de celles qui dépendent du modèle spatio-temporel.

6.2 Résultats de détection

Les résultats de détections annoncés dans le chapitre précédent (5.4.2) révèlent que les (VP) détectés concernent des référencements multi-linéaires, réalisés par la main droite, le regard et la tête. Ceci confirme la déduction à laquelle nous avons abouti dans l'analyse de combinaisons gestuelles réalisées à partir du corpus *« SignCom »* (Voir conclusion du chapitre 4). Nous avons résumé les taux de détection selon la composition corporelle réalisant le référencement détecté dans le tableau (6.1).

Vidéo \ Composition	R et SL(%)	R et PT index(%)
Vidéo2	50	50
Vidéo3	50	50

TABLE *6.1 – Taux de détection de vrai positifs par type de référencement.* ***R et SL*** *: Regard et Signe localisé.* ***R et PT index*** *: Regard et PoinTé par l'index de la main droite*

Nous avons constaté que les référencements non détectés ont la forme : 1) d'un signe locatif réalisé par la main gauche ou 2) d'un signe de pointé très bref réalisé par l'index de la main droite (main immobile).

Afin de détecter les raisons pour lesquelles nous avons obtenu des faux positifs, nous allons remettre en question les paramètres caractéristiques du référencement : la vitesse de la main droite et la distance entre main droite et locus, entre tête et locus.

Les décalages présents entre certains vrais référencements détectés et référencements annotés (vérité terrain) nous permettent d'émettre l'hypothèse que certains comportements ne sont pas représentés par le modèle 2D.

6.3 Entrées / Sorties

Les vidéos d'énoncés signés sont les entrées du système de reconnaissance. Leur enregistrement a été réalisé sou plusieurs conditions. Nous allons établir le bilan des limites engendrées par les conditions d'enregistrement des corpus vidéo. Ceci influence les résultats de détection de régions d'intérêt. En conséquence, l'interprétation du mouvement des ROIs (en $2D$) est biaisée et donc la classification des paramètres mesurés c'est-à dire les vitesses instantanées et les distances l'est également.

6.3.1 Acquisitions du corpus

Afin d'évaluer la capacité de l'algorithme à gérer les problèmes liés au contenu des vidéos, nous prendrons en considération les conditions dans lesquelles ont été réalisées les sessions d'enregistrement, à savoir la caméra statique, le fond fixe et uniforme, les vêtements noirs que porte le signeur qui reste seul dans l'espace de travail et est constamment filmé de face. Ensuite, nous évaluerons la fiabilité de l'hypothèse de la lenteur des mouvements de la tête.

L'absence de données du regard (coordonnées de la cible) rapportés dans le tableau (3.2) est gérée par le programme de reconnaissance car celui-ci gère l'absence de données du regard d'une manière transparente à l'utilisateur moyennant des tests conditionnels.

Le fort contraste que produit le fond uniforme et les habits du signeur avec la couleur de peau facilitent la détection de pixels appartenant à la zone peau. Cependant, ce critère impose des contraintes sur l'habillement et sur la couleur de peau du signeur. le signeur est contraint de porter des vêtements à manches longues et ras de cou pour une extraction optimale des régions d'intérêt.

Nous nous sommes confrontés au problème de non respect de cette contrainte qui n'a pas été sans conséquences sur la segmentation de la ROI Main droite et celle de la tête. L'image (6.1 - a) illustre une imprécision dans l'estimation de la ROI de la tête due à une encolure large du vêtement porté par le participant. Ceci peut biaiser les mesures du mouvement de la tête dans le sens où cette mesure représente la partie inférieure de la ROI tête et donc demeure immobile pendant que le tête bouge. Pour cela, nous avons donné un coefficient important à la zone supérieure de la ROI tête. Afin d'évaluer l'exactitude des mesures de déplacement moyennant cette méthode, sachant que nous ne pouvons le faire automatiquement, nous avons relevé quelques mesures de déplacement de la ROI tête et vérifié leur cohérence avec les mouvements annotés à partir de l'observation de la vidéo. Nous avons observé, dans les vidéos, des mouvements non reportés par les mesures de déplacement de ROI de la tête. L'image (6.1 - b) montre une estimation imprécise des ROI de la main droite.

Nous remettons en cause l'imprécision de détection de la main droite qui représente une des causes de la détection de faux positifs, parallèlement avec d'autres facteurs que nous essaierons de déterminer au fur et à mesure dans ce chapitre.

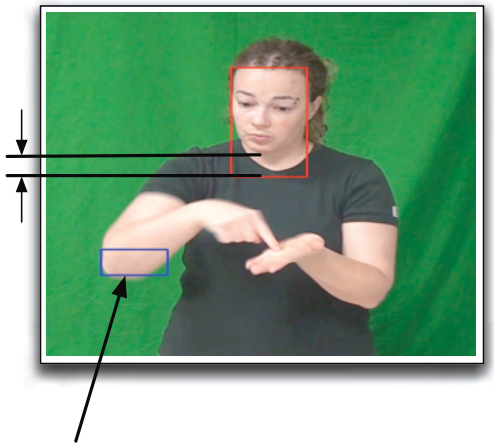


FIGURE *6.1 – Segmentation imprécise de la ROI : Cadrage imprécis de la Tête (encolure large) et mauvaise estimation de la position de la main droite (manche courte)*

6.3.2 Modèles spatio-temporel

Le modèle gestuel de référencement mis en jeu dans le processus de reconnaissance est de dimension 2 et représente la transformation du modèle tri-dimensionnel de gestes de référencement (Voir chapitre (4)). Dans les deux sections suivantes, nous étudierons les limites du modèle 2D ainsi déduit.

6.3.2.1 Modèle 2D

Le modèle 2D mis en jeu dans le processus de référencement consiste en deux paramètres : 1) la variation de la distance entre ROIs de la tête et de la main droite et 2) Le profil d'évolution de la vitesse instantanée de déplacement de ROIs. L'implémentation du modèle 2D est simple et ne dépend pas de la morphologie du signeur. Les paramètres pris en compte dans le modèle le rendent représentatif des comportements de la main droite et de la tête lors d'un référencement en terme de distance $2D$ entre ROI tête et ROI main droite.

Nous avons fait appel à la méthode de calcul des extremums locaux afin de détecter les moments clés selon les pics de distances entre ROIs. La détermination des extremums locaux nécessite, en premier lieu, un filtrage sélectif afin d'éliminer les valeurs aberrantes dues aux erreurs d'estimation de positions de ROIs. Nous avons relevé les pics de vitesse dans le but d'étudier l'évolution de la vitesse de la tête et de valider la stratégie de segmentation temporelle de séquences de référencement selon l'enchaînement (– changement de la direction du regard – fixation de l'espace de signation) sans faire appel aux coordonnées de la cible du regard *« FaceLab »*, dont le taux de perte est important (Voir tableau qui résume les taux de perte de données (3.2) et la caractérisation de ROIs dans 5.4.1.3). Une vérification approximative est possible en comparant les

positions des extremums dans le temps aux mêmes moments que les fixations du regard observés dans la vidéo. Cependant, nous n'avons pas pu confronter les séquences de vidéo segmentées selon la stratégie des extremums locaux avec la vérité terrain afin de quantifier l'imprécision et ceci à cause de mesures manquantes dans les données de la cible du regard.

L'évolution de la vitesse 2D de la ROI main droite relevée dans la session 2 est illustrée dans (6.2). Nous avons réalisé un filtrage sélectif (3 itérations) afin d'éliminer les données aberrantes. Dans la figure (6.2), nous constatons deux motifs identiques séparés par un palier. Après comparaison manuelle, les extremums locaux, marqués par des disques, ne sont contenus dans aucun intervalle de référencement.

D'autre part, nous avons considéré que l'évolution de la vitesse instantanée de la main droite est un paramètre nécessaire mais pas suffisant pour caractériser un référencement manuel (Voir explication dans 4.4.3.1), et de même pour celle de la tête. De ce fait, nous avons intégré ce paramètre dans le modèle spatio-temporel et donc pris en considération lors de la reconnaissance de référencement. Par ceci nous nous fixons comme objectif d'améliorer le résultat de classification jugé faible en fusionnant les résultats d'estimation de coefficients de similarité entre mesures $2D$ et modèle 2D. Concrètement, l'objectif serait de calculer le produit des coefficients d'appartenance à chaque classe ($C1$ et $C2$) afin de fournir une décision finale pondérée.

6.3.2.2 Modèle temporel

D'un autre côté, nous avons pris en considération l'ordre d'enchaînement des gestes de référencement. Nous nous sommes inspirés des modèles linguistiques de référencement afin de construire un modèle temporel dont les paramètres ont été ajustés moyennant l'analyse du corpus *« Websourd »* présenté dans le chapitre (3). Nous avons abouti à plusieurs types d'enchaînement gestuel et par conséquent nous avons choisi un modèle temporel non rigide qui décrit les relations temporelles, observées dans le corpus, selon la logique d'Allen (Voir 4.2). Cependant, le modèle résultant ne tient compte que de l'orientation du regard qui précède les mouvements de la main et de la tête.

6.3.3 Les sorties

Dans cette section, nous passerons en revue les sorties fournies par chaque étape de l'algorithme de reconnaissance.

Les données de sortie $2D$ fournies sont classées selon une combinaison de l'ensemble de la classe $C2$.

La validation de la décision finale, qui représente la sortie du système de reconnaissance, nécessite une confrontation entre la valeur de la décision (référencement ou pas) et la vérité terrain. Pour cela, nous avons fait appel à un expert en (LSF) afin d'évaluer l'exactitude des vrais positifs (VP) et d'ajuster la délimitation des séquences vidéos détectées en tant que (VP) en décalage par rapport à la vérité terrain.

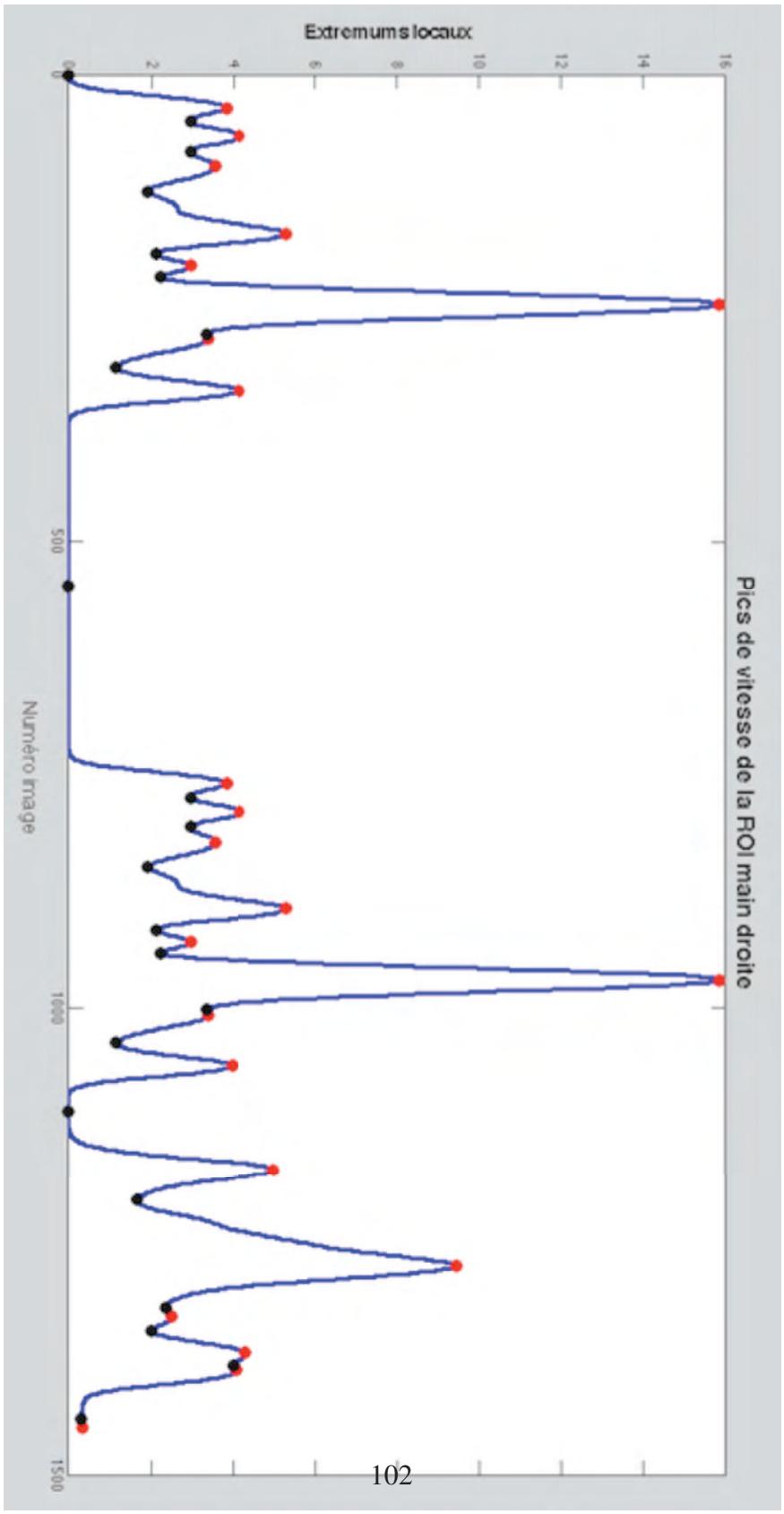


FIGURE 6.2 – *Les pics de la vitesse 2D de la ROI main droite correspondant à la session 2*

6.4 Etapes de l'algorithme

6.4.1 Détermination des ROI de peau

Nous avons utilisé plusieurs hypothèses afin de segmenter les ROIs et de déterminer sa nature (tête ou main droite) :

- Un intervalle de valeurs de couleur de pixels correspondants à celles de la texture de la peau.
- Un découpage de l'image en zones clés (Voir figure 5.4) traduisant les zones de la tête et de la main droite.
- La faible amplitude de la vitesse d'un mouvement réalisé par la tête.

L'intervalle de valeurs de couleur de pixels a été appliqué sur tous les signeurs du corpus *« SignCom »* et a permis une segmentation satisfaisante (Voir figures 5.6). Toutefois, le découpage de l'image en zones clés a imposé une géométrie particulière au cadrage de la ROI (en rectangles ce qui a mené en partie au résultat illustré dans les figures (6.1 - a) et (6.1 - b) et qui consiste en un cadrage peu précis. De plus, ce découpage ne tient pas compte de la main quand celle-ci se trouve au-dessus de la tête du signeur. Nous notons que l'intervalle de valeurs de pixels de peau ne peut permettre la détection de ROI dans le cas d'un signeur dont la couleur de peau est foncée. Au delà de la valeurs de couleur de pixels de peau, ce cas peut même entraîner le changement des conditions d'acquisition, à savoir les vêtements et le fond. Toutefois, nous avons pu détecté les positions de la main droite et de la tête à des taux variés (voir tableau 6.2) que nous jugeons, en moyenne, satisfaisants pour la zone main droite mais insuffisant pour la tête à cause du biais engindré par l'algorithme de segmentation et le non respect des certaines conditions d'enregistrement (l'habit recommandé). Nous précisons que nous avons utilisé les sessions d'enregistrement qui ont le meilleur taux de détection des ROIs.

Taux de détection / Vidéo	Tête(%)	Main droite(%)
Vidéo2	35.4	90
Vidéo3	48.63	75.57

TABLE *6.2 – Taux de détection des zones peau correspondantes à la tête et à la main droite*

Nous notons que la méthode de détermination de ROI peau a nécessité un calibrage peu fréquent pour la tête en tenant compte de l'hypothèse de la faible amplitude du mouvement de la tête. Nous avons, par contre, réalisé un calibrage plus fréquent pour la main droite. Nous avons remis en question la première hypothèse (mouvement de la tête)

et testé la détermination de zones ROI peau avec un calibrage plus fréquent de la tête et nous n'avons enregistré aucune amélioration. Ceci montre que l'hypothèse en question est vérifiée dans le corpus *« SignCom »*, sur les vidéos de calibrage précisément (Voir tableau 6.2).

6.4.2 Estimation des mouvements des ROIs

Cette tâche a pour objectif de vérifier l'hypothèse que la distance permet de compenser la donnée de profondeur manquante au modèle spatial projeté. Elle consiste à déterminer une métrique qui exprime :

- le déplacement bi-dimensionnel des ROIs (vue de face) correspondants à la main droite et à celle de la tête.
- l'évolution de la distance $2D$ entre ROIs.

Les mesures de la distance relative nous a permis de déduire une stratégie de repérage de certains segment de référencement et ceci moyennant le calcul des positions d'extremums locaux. Nous avons relevé, dans la session (2) [1] un taux total de 31% de segments de référencement dont les distances relatives : ROI Main droite – ROI Tête présentent au moins un extremum local.

Ainsi, la mesure de la distance uniquement ne permet pas de compenser la donnée de profondeur manquante.

6.4.3 Mesures et classification

Les règles de déduction dépendent du modèle 2D et des mesures obtenues. Pour le modèle spatial, la comparaison consiste à déterminer si la variation des distances entre ROI et l'amplitude des vitesses instantanées de la main droite sont régies par les mêmes lois de distributions décrites par le modèle 2D proposé.

Pour l'aspect temporel, la comparaison consiste à déterminer si le décalage temporel entre les mouvements de la tête et de la main droite correspond à celui décrit par le modèle temporel.

Au delà de cette comparaison, juger de la fiabilité de la décision binaire est une tâche qui met en jeu 1) la méthode de déduction de conformité entre modèles et mesures et 2) la validité du modèle 2D. La validité du modèle 2D dépend de celle de la méthode de projection – $3D \rightarrow 2D$ – et du modèle spatial tri-dimensionnel proposé dans le chapitre (4).

Méthode de projection ($3D \rightarrow 2D$) **:** Nous avons utilisé l'appariement entre paramètres du modèle $3D$ et le modèle 2D $2D$. Nous avons relevé les profils de variation de la distance entre ROIs décrits par les classes $C1$ et $C2$. Nous avons remarqué que seule la classe $C2$ inclut des (VP). Nous résumons ceci dans le tableau (6.3).

Nous obtenons ainsi une détection de 65% des segments de référencement présents dans deux sessions d'enregistrement d'une durée totale de 108 secondes. Les segments

1. D'une durée d'une minute environ

Vidéo \ Taux de détection	VP(%)	FP(%)
Vidéo2	80	20
Vidéo3	50	50

TABLE *6.3 – Taux des segments détectés appartenant à la classe C2. Les (VP) sont les référencements détectés correspondants à la vérité terrain. Les (FP) sont les référencements détectés ne correspondant pas à la vérité terrain.*

détectés sont considérés, d'après la classification, des pointés à amplitude faible (proches du signeur). Moyennant l'identification de deux classes de mesure de pointé (proche et loin du signeur), le reprérage des extremums locaux se révèle une méthode pertinente pour la détection de pointés manuels.

6.4.4 Décision

Afin de pondérer l'appartenance d'un segment vidéo à une structure de référencement, nous avons attribué des coefficients aux segments selon le nombre de classes auxquelles ils appartiennent. Dans le tableau (6.4), nous classons les segments détectés selon le nombre de classes auxquelles ils appartiennent. Le graphe (6.3) traduit la possibilité de réduction de (FP) utilisant ce critère.

Nous avons vu que la déduction de conformité entre modèle tri-dimensionnel et coordonnées spatiales ou entre modèle temporel et décalages (entre mouvements) dépend du taux de détection obtenu. Afin de rendre la décision finale de reconnaissance robuste aux erreurs d'estimation, nous avons considéré que la décision finale est le résultat du produit des indices de conformité des modèles de distance et de vitesse $2D$.

	Vidéo \ Classe et taux d'appartenance (%)	1	2	3
(VP)	Vidéo2	62	19	6
	Vidéo3	60	40	0
(FP)	Vidéo2	95	12	0
	Vidéo3	92	28	0

TABLE *6.4 – Taux de segments (VP) et (FP) appartenant à {1 ;2 ;3} classe(s)*

Les graphes (6.3 - a & b) révèlent que les taux de (VP) diminue de 5% plus que celui de (FP) dans la session 2 et de 10% dans la session 3. En effet, nous observons dans les

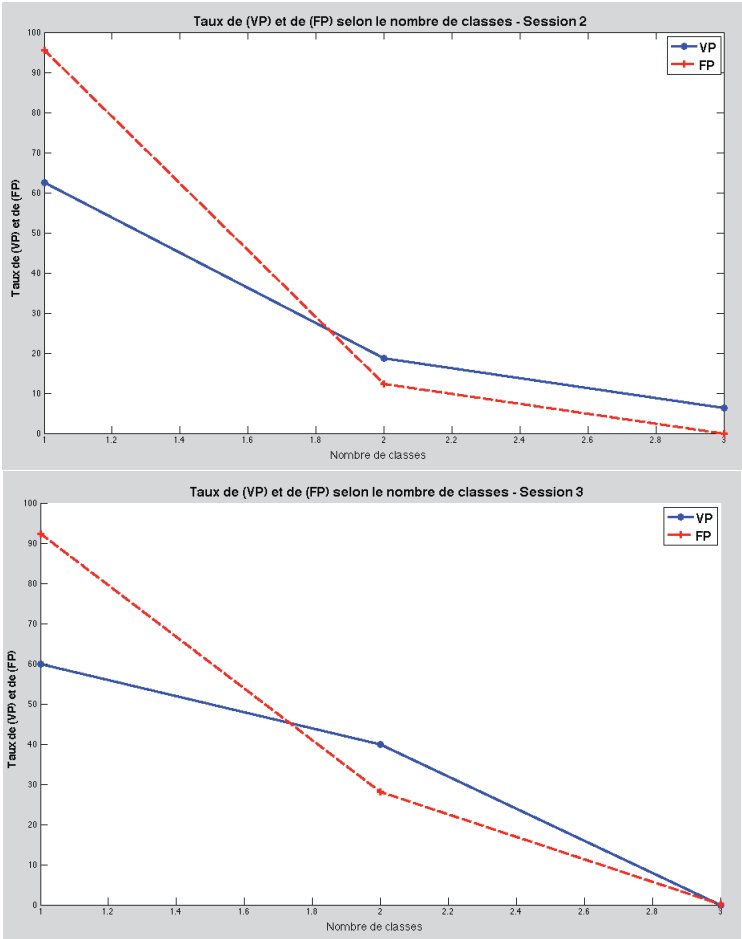


FIGURE *6.3 – Les taux de (VP) et de (FP) selon le nombre de combinaisons de classe pour a) la session 2 et b) la session 3*

deux graphes (a & b) que les taux de (VP) et (FP) pour la classe de combinaisons de degré (2) est inférieure à celle de la classe de combinaisons de degré (1). D'une part, nous constatons ainsi que le taux de (FP) a diminué. D'autre part, nous enregistrons un taux de réduction du taux de (VP) inférieur à celui de (FP). Ainsi, la classification des combinaisons par le nombre de gestes représente donc une piste exploitable pour la réduction du taux de faux positifs détectés.

6.5 Bilan

Dans cette section, nous énumérerons les points positifs et négatifs de l'algorithme de reconnaissance réalisé puis pointerons les apports et les limites de la méthodologie de reconnaissance retenue.

6.5.1 Algorithme de reconnaissance

Nous avons mis en place un algorithme simple du point de vue coût d'implémentation et une architecture modulaire. Toutefois, nous n'avons pas automatisé certaines étapes qui sont : la mise à jour du modèle $2D$ et de l'historique de l'enchaînement gestuel de référencement rencontré à un instant antérieur de la vidéo. Le programme est simple, en particulier le module de détection de zones peau qui est fortement lié aux conditions d'acquisition des enregistrements (fond uniforme, caméra fixe, etc.).

Nous notons que la sortie du programme se présente sous forme de segments de vidéos étiquetés (VP) ou (FP) mais ne fournit pas de détails sur le type de référencement en question (Signe localisé, signe de pointé par l'index, etc.).

Compte tenu des résultats de détection, nous expliquons le fait que, même avec une segmentation imprécise de la ROI main droite, le signe de pointé manuel a été détecté par l'hypothèse que la variation de la distance mesurée entre la ROI main droite et la ROI tête permet de caractériser le référencement.

Le coût d'exécution est long car le module de caractérisation des mouvements $2D$ des ROIs requiert le stockage temporaire du contenu des valeurs et positions des pixels des ROIs tête et main droite.

6.5.2 Méthodologie de reconnaissance

Le raisonnement qui a permis la mise en place d'une architecture logicielle de détection d'événements de référencement prend en entrée la description du contenu des images enregistrées par une caméra monoculaire en tournage intérieur. La caractérisation de données d'entrée ne dépend pas de la morphologie du signeur et tient compte le fait que les images peuvent être de mauvaise qualité. En effet, les métriques sur lesquelles se basent, en partie, la décision finale (distance et vitesses) représentent la moyenne de mesures effectuées sur une suite d'images.

Nous avons, d'une part, décrit l'aspect spatial des données d'entrée à travers des

modèles gestuels tri-dimensionnels et en déduit un modèle bi-dimensionnel réalisable. Nous avons quantifié et formalisé l'aspect temporel des réalisations gestuelles via la logique temporelle d'Allen.

L'annotation des corpus utilisés a été d'une grande aide dans l'affinement du modèle spatio-temporel $3D$. Cependant, la phase de modélisation a nécessité un nombre important de vidéos enregistrées et par conséquent un temps considérable pour leur annotation. L'exploitation du modèle $3D$ sous sa forme réduite $2D$ a nécessité aussi une étape de calibrage qui s'est révélée intéressante pour repérer les référencements par le regard ou par la main mais ne permettent pas de repérer les pointés par le tête.

6.6 Conclusion

Dans ce chapitre, nous avons identifié les limites du système de reconnaissance et proposé des pistes d'amélioration des résultats obtenus.

Au niveau des données fournies en entrée, l'application du protocole d'acquisition est nécessaire pour garantir des mesures précises. Les résultats fournis par le système de reconnaissance ne fournit pas de détails sur la fonction linguistique correspondante au référencement en question. Les étapes de l'algorithme de reconnaissance ont permis de confirmer ou d'infirmer les hypothèses du faible déplacement de la tête et du paramètre caractérisant le référencement en $2D$. Nous avons identifié une stratégie de repérage de certains segment de référencement et les deux classes de mesure de signes de pointés manuel.

D'autre part, nous avons expliqué, dans ce chapitre, la détection de faux références-ments et proposé des solutions pour diminuer le taux de (FP) correspondant. Toutefois, cette solution se conjugue avec la diminution du taux de vrais référencements détectés (VP). Nous avons proposé une stratégie qui permet de remédier à la diminution de (VP) en classifiant les combinaisons de geste par degré. Cette conclusion s'avère pertinente car elle renseigne sur la diminution du taux de (VP) dans une perspective d'amélioration du résultat de détection de référencement.

D'une manière générale, notre méthodologie de reconnaissance a permis de déterminer des caractéristiques du référencement qui sont indépendantes de la morphologie du signeur. La phase de calibrage s'est avérée pertinente pour repérer les référencements par le regard et / ou la main.

Chapitre 7

Perspectives et Conclusions

Sommaire

7.1 Introduction

Notre travail a porté sur le regard comme acteur important dans la compréhension d'un discours en langue des signes. Notre objectif de départ était d'extraire les éléments qui, à la fois, structurent un énoncé signé et sont liés au rôle du regard. Comme éléments de départ, nous nous sommes basés sur l'espace de signation et les gestes. De ce fait, nous nous sommes intéressés aux fonctions linguistiques qui présentent une relation entre le regard et l'espace de signation telles que le signe localisé, le signe en mouvement, le signe de pointé et le signe locatif. Nous avons choisi de qualifier la fonction commune à ces notions par *« référencement »*.

L'objectif final est de mettre en place un système de reconnaissance d'un ou de plusieurs formes de référencement. Pour cela, nous avons proposé une description fine des réalisations gestuelles de référencement provenant de corpus variés. Le résultat de la description est une représentation géométrique tri-dimensionnelle car nous avons constaté que la caractérisation du rôle du regard et de l'espace de signation est liée à l'algorithme de reconnaissance mis en place tient compte d'un modèle bi-dimensionnel qui représente la transformation $3D \rightarrow 2D$ du modèle description de référencement.

7.2 Réalisations

7.2.1 Corpus

Nous avons utilisé des corpus variés en termes de format de données (vidéos annotées, capture de mouvement et du regard). Le choix de format s'est basé sur l'objectif de

chaque étape de la méthode de reconnaissance. En premier lieu, nous avons besoin d'un corpus riche en données pour pouvoir décrire finement le référencement. L'objectif de l'utilisation d'un second corpus était de tester le modèle descriptif de référencement sans fournir au programme de test des informations sur l'orientation du regard. Seul le troisième corpus qui était uniquement en format vidéo. L'objectif était de se mettre dans un contexte de discours signé et de vérifier si on peut adapter le modèle descriptif à l'identification du référencement dans une vidéo. Les conditions d'enregistrement ont été établies dans le but de rendre optimal les résultats de segmentation spatiale et temporelles. Pour cela, les protocoles d'acquisition sont différents pour chaque corpus.

La mise en oeuvre de formats de données variés d'une même session d'enregistrement a nécessité des étapes de mise en correspondance entre formats en termes de synchronisation temporelle et projection spatiale. Cette étape a été automatisée grâce à un logiciel de synchronisation. La vérification du bon déroulement de la synchronisation était coûteuse en termes de temps.

7.2.2 Modélisation

Le modèle descriptif du signe de pointé inclut trois mesures :
- Le décalage temporel entre gestes (inter-gestuel).
- La position spatiale de la composante corporelle.
- La vitesse de la main droite.

Les mesures réalisées sur les corpus de modélisation se basent sur 1) des annotations manuelles des mouvements des mains, de la tête et de l'orientation du regard et sur 2) des représentations géométriques tri-dimensionnelles des mains, de la tête et de la cible du regard.

La mesure de décalages inter-gestuel a fait apparaître plusieurs motifs gestuels. La différence entre les motifs gestuels concerne l'ordre chronologique des gestes et le nombre de composants corporelles impliqués (degré de combinaison). Une réflexion menée sur la simplifications du nombre de motifs gestuel nous a conduit à utiliser la logique temporelle d'Allen et le principe de transitivité.

La mesure de profil de vitesse de la main droite a fait apparaître plusieurs types de profils. Nous avons pris en considération les profils les plus fréquents.

La mesure de distances relatives a montré que la main droite et la tête marquent l'intérieur de la frontière estimée du locus alors que le regard marquent une zone approximative.

Ces mesures nous ont permis de construire trois types de modèles du référencement (temporel, spatial et dynamique). Les règles déduites de ces modèles ont été utilisées dans le système de reconnaissance.

7.2.3 Programmes de reconnaissance

Nous avons ainsi décrit les aspects spatial, temporel et dynamique du référencement en se basant sur des gestes d'un signeur dans un discours en langue des signes. L'objectif

suivant est d'utiliser ces modèles afin de reconnaître un référencement dans une vidéo en langue des signes. Rappelons que le référencement est un concept présent dans plusieurs fonctions linguistiques dont 1) le signe localisé, 2) le signe de pointé, 3) le signe locatif et 4) le signe en mouvement abordées dans le chapitre (2). Le système de reconnaissance prendra en entrée des modèles descriptifs du référencement et une vidéo. Les données d'entrée appartenant à des dimensions différentes $3D$ et $2D$, nous avons choisi de transformer le modèle descriptif $3D$ en un modèle $2D$ et ceci en éliminant la composante de profondeur.

L'algorithme de reconnaissance mis en place se base sur deux approches : ascendante et descendante du fait que les données d'entrée appartiennent à deux niveaux différents : bas niveau (pixel) et haut niveau (description de gestes). D'un côté, l'algorithme de reconnaissance permet d'interpréter les pixels en régions d'intérêt (regroupement de pixels) puis en mouvements (déplacement des ROI). De l'autre côté, l'algorithme de reconnaissance permet de fusionner les modèles descriptifs spatial et temporel.

7.2.4 Evaluation des résultats

Les résultats de l'algorithme de reconnaissance se présentent sous deux formes : 1) Une classification des mesures réalisées sur le corpus test et 2) Une décision finale sur l'appartenance des classes en question au modèle de référencement. Nous avons obtenu un taux de détection satisfaisant (65%).

Ainsi, la combinaison des paramètres de la vitesse et de la distance relative permet de caractériser le référencement avec :

- des contraintes d'acquisition de corpus.
- un choix de méthode de projection $2D$ à partir d'une représentation géométrique $3D$.
- des hypothèses émises sur les mouvements du signeur.

Cependant, nous avons constaté que le choix de passage $2D \rightarrow 3D$ c'est-à dire l'élimination de la composante de profondeur a entraîné des erreurs dans le modèle de distance relative Main - Tête. D'autre part, l'exactitude des résultats fournis par chaque étape de l'algorithme de reconnaissance dépendent du protocole d'enregistrement mis en place.

La méthodologie de reconnaissance adoptée est réutilisable pour la modélisation et la reconnaissance d'autres types de structures syntaxiques faisant intervenir les gestes manuels et/ou non-manuels.

Toutefois, la méthodologie de reconnaissance requiert des contraintes dans l'acquisition de corpus pour simplifier la complexité de l'analyse des images $2D$ de vidéos d'énoncés de langue des signes.

D'autre part, nous avons traité les cas de référencement unique et de référencement vers des loci localisés d'une manière particulière.

7.3 Perspectives

Les modèles $3D$ construits décrivent les profils dynamiques de la main droite et la distance locus – composante corporelle. Afin de fusionner ces modèles pour les rendre représentatif de la vérité terrain, on pourrait prévoir une combinaison de ces deux paramètres.

La prise en compte de plusieurs vues d'une même vidéo serait une solution au problème d'adaptation entre modèle $3D$ et données $2D$ d'entrée fournies au système de reconnaissance. Pour cela, une décomposition du modèle $3D$ en deux modèles $2D$ seraient plus appropriées.

Ainsi, il serait intéressant de produire des corpus vidéos avec deux vues (de face et de profil). Donc, envisager une automatisation de la synchronisation spatiale et temporelle en amont simplifiera le traitement des données $2D$.

D'une manière générale, la méthodologie de reconnaissance mise en place permettra d'analyser les cas de référencement non traités dans cette étude. Afin d'analyser les référencements multiples à condition d'exclure les gestes non-manuels car, manuellement, il relève de l'impossible de pouvoir délimiter les signes de pointé correspondants.

7.4 Conclusion générale

Nous avons déduit que la prise en compte de plusieurs vues devrait être incluse dans le cahier de charge d'acquisition de corpus. Ceci montre que la phase de modélisation joue un rôle dans la détermination du cahier de charge de construction de corpus.

Ainsi, il serait intéressant de composer la phase de modélisation en deux parties : avant et après enregistrement de corpus. La phase avant-enregistrement de corpus se basera sur les modèles gestuels de fonctions linguistiques.

Dans cette étude, nous nous sommes concentrés sur le référencement en tant que concept contenu dans plusieurs fonctions linguistiques. Les résultats fournis par le système de reconnaissance concernent les fonctions de pointé.

Afin de réutiliser le système de reconnaissance pour caractériser les fonctions de *« signes en mouvement »* et *« signes localisés »* qui font intervenir la signification du signe réalisé, il serait nécessaire d'introduire le sens des gloses dans la phase de modélisation du référencement.

Cependant, la reconnaissance de fonctions de référencement faisant intervenir le sens des gestes rajoutera une problématique supplémentaire à celle de reconnaissance de référencement basé uniquement sur les gestes. De ce fait, la fusion de cette étude avec celle effectuée sur la reconnaissance de configuration de signes pourrait apporter des éléments de réponse à cette problématique.

ANNEXE A

Comparatifs de méthodes de capture du regard

A.1 Détection de l'oeil

– Il existe plusieurs méthodes qui satisfont plus ou moins nos critères [Hansen 2010].
– Nous en avons choisi trois [[Hansen 2005] (1), [Xu 2009] (2) et [Asano 2011] (3)] (voir tableau A.1)

TABLE *A.1 – Tableau comparatif des méthodes de détection de l'oeil*

Critères / Méthodes	(2)	(3a)	(3b)	(3c)	(3d)	4	5	6	7	8
(1)	✓	✓	✓	✓	✓	✓	✓	✓	-	✓
(2)	✓	✓	✓	-	-	-	✓	✓	✓	✓
(3)	✓	✓	-	-	-	-	✓	✓	-	✓

La méthode (2) ne satisfait pas les critères (3) c, (3) d et (4), mais, elle est simple à mettre en oeuvre.

A.2 Détection de l'orientation du regard

Les objectifs de l'analyse :

1. Estimer la direction du regard
2. Déterminer la cible pointée par le regard (la main, l'espace, autre)

[Shih 2000] a montré qu'il faut deux caméras et deux sources lumineuses au minimum pour avoir un suivi robuste de la direction du regard même quand la personne filmée bouge sa tête.
[Valenti 2010] montre que les mouvements de la têtes introduisent des erreurs plus importantes sur les mesures d'angles du regard.
– Il existe plusieurs méthodes qui satisfassent plus ou moins nos critères.

TABLE *A.2 – Tableau comparatif des méthodes de détection de la direction du regard*

Critères / Méthodes	(2)	(3a)	(3b)	(3c)	(3d)	4	6	7	8
[Beymer 2003]	[1]	-	-	-	-	-	±0.6°	glint	✓
[Brolly 2004]	[2]	-	-	-	-	-	±0.8°	glint	✓
[Shih 2004]	30fps	✓(c+)	-	-	-	-	<1°	glint	✓
[Coutinho 2006]	30fps	✓	-	-	-	✓	±1°	glint	(5sl)
[Wallhoff 2006]	25fps	±15°	-	-	-	-	±4.15°	AAM	(1c)
[**?**]	-	>20°	-	-	-	✓	✓	AAM	(1c)
[Takatani 2010]	✓	±20°	-	-	-	-	±2.7°/h	AAM	✓
[Valenti 2012]	✓	>20°	-	-	-	-	±2°	AAM	✓

– Nous en avons choisi dix ((voir tableau A.2).

Même si la méthode de détection de l'Iris par réflexion de la lumière infra-rouge semble directe et plus rapide, la méthode de modèles d'apparence semble la moins coûteuse en terme de mise en oeuvre et surtout la plus robuste aux changements de positions de la tête. Cependant, il reste à déterminer quelle algorithme choisir.

Le tableau comparatif (A.3) établit le bilan des résultats des méthodes les plus précises. La méthode [Lu 2011] reste la plus précise même si elle requiert un calibrage

TABLE *A.3 – Tableau comparatif des méthodes de détection de la direction du regard basée sur les modèles d'apparence*

Critères / Méthodes	Précision (°)	Tête(°)	Modèle oeil	Multi-vues	Occultations
[Lu 2011]	2.38 ou 6 [3]	free	-	[4]	-
[Nakamatsu 2012]	6 [5]	-	3D-AAM	-	-
[Valenti 2012]	≈4.6 [6]	±30	-	[7]	-
[Salam 2012]	nc	-	-	-	-

des positions de l'oeil et de la tête.

1. fréquence d'échantillonnage de 20 images par seconde, largement inférieure à 60 images par seconde, la fréquence d'échantillonnage des capteurs de mouvement
2. c+ :avec calibrage des mouvements de la tête
3. sans calibrage
4. calibrage : varier les fixations puis les rotations de la tête
5. horizontal et vertical. Le regard de ±15°

6. ≈8° sans calibrage
7. avec calibrage et retargeting

Bibliographie

[Al-Jarrah 2001] O Al-Jarrah et Alaa Halawani. *Recognition of gestures in Arabic sign language using neuro-fuzzy systems*. Artificial Intelligence, vol. 133, pages 117–138, 2001. (Cité en page 79.)

[Allen 1983] James F. Allen. *Maintaining knowledge about temporal intervals*. Communications of the ACM, vol. 26, no. 11, pages 832–843, Novembre 1983. (Cité en pages 50 et 52.)

[Allen 1994] JF Allen et G Ferguson. *Actions and events in interval temporal logic*. Journal of logic and computation, 1994. (Cité en page 49.)

[Asano 2011] Masayuki Asano, Hironobu Takano et Kiyomi Nakamura. *Eye detection method robust to facial pose changes for eye input device*. 2011 IEEE International Conference on Systems, Man, and Cybernetics, vol. 1, pages 602–607, Octobre 2011. (Cité en page i.)

[Badaloni 2000] S Badaloni et M Giacomin. *A fuzzy extension of Allen's interval algebra*. AI* IA 99 : Advances in Artificial Intelligence, vol. 2, no. 1, 2000. (Cité en page 82.)

[Bedregal 2006] B. R. C. Bedregal, G.P. Dimuro et A.C.R. Costa. *Fuzzy Rule-Based Hand Gesture Recognition for Sign Language Processing*. In International Joint Conferences IBERAMI/SBIA/SBRN, Workshop on Computational Intelligence, Ribeirão Preto, 2006. WCI. (Cité en pages 79 et 82.)

[Beymer 2003] D. Beymer et M. Flickner. *Eye gaze tracking using an active stereo head*. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., volume 2, pages II–451–8. IEEE Comput. Soc, 2003. (Cité en page ii.)

[Bras 2004] G Bras, A Millet et A Risler. *Anaphore et déixis en LSF Tentative d'inventaire des procédés*. Rapport technique, Résumé du Colloque Linguistique de la LSF : recherches actuelles, Université Lille 3, 2004. (Cité en pages 13 et 15.)

[Brolly 2004] X.L.C. Brolly et J.B. Mulligan. *Implicit Calibration of a Remote Gaze Tracker*. 2004 Conference on Computer Vision and Pattern Recognition Workshop, pages 134–134, 2004. (Cité en page ii.)

[Chetelat 2010] E. Chetelat. *Les gestes non-manuels en langue des signes française, annotation, analyse et formalisation : Application aux mouvements de sourcils et aux clignements des yeux*. PhD thesis, Université d'Aix Marseille, 2010. (Cité en page 43.)

[Coutinho 2006] Flavio Coutinho et Carlos Morimoto. *Free head motion eye gaze tracking using a single camera and multiple light sources*. In 2006 19th Brazilian Symposium on Computer Graphics and Image Processing, pages 171–178. IEEE, Octobre 2006. (Cité en page ii.)

[Cuxac 2000] Christian Cuxac. Faits de Langues - La langue des signes française (LSF) - Les voies de l'iconicité. Faits Des Langues : Ophrys, Paris, 2000. (Cité en pages 2, 10 et 16.)

[Cuxac 2003] C. Cuxac. *Une langue moins marquée comme analyseur langagier : l'exemple de la LSF : Langue des signes française (LSF) : Enjeux culturels et pédagogiques*. La Nouvelle revue de l'adaptation et de la scolarisation, vol. 1, no. 23, pages 19–30, 2003. (Cité en page 10.)

[Cuxac 2005] Christian Cuxac. *Les langues des signes : analyseurs de la faculté de langage*, 2005. (Cité en pages 16 et 17.)

[Dalle 2009] Patrice Dalle et François Lefebvre-albaret. *Analyse des pointages en LSF par traitement automatique de vidéos.* In Symposium "Du geste au signe : le pointage dans les langues orales et signées - Université de Lille", 2009. (Cité en pages 48, 70 et 72.)

[Deuchar 1984] M. Deuchar. British sign language. Croom Helm., London, 1984. (Cité en page 14.)

[Elliott 2007] R. Elliott, J. R. W. Glauert, J. R. Kennaway, I. Marshall et E. Safar. *Linguistic modelling and language-processing technologies for Avatar-based sign language presentation*. Universal Access in the Information Society, vol. 6, no. 4, pages 375–391, Octobre 2007. (Cité en page 25.)

[Engberg-Pedersen 1986] E Engberg-Pedersen. *The use of space with verbs in Danish Sign Language*. Sign Language Research. Amsterdam, 1986. (Cité en page 14.)

[Engberg-Pedersen 2003] E. Engberg-Pedersen. *From pointing to reference and predication : pointing signs, eyegaze, and head and body orientation in Danish Sign Language*. Movie, pages 269–292, 2003. (Cité en page 16.)

[F. Lefebvre-Albaret 2010] P Dalle F. Lefebvre-Albaret et F. Lefebvre-Albaret .P. Dalle. *Requête dans une vidéo en Langue des Signes Modélisation et comparaison de signes*. RFIA, 2010. (Cité en page 43.)

[Fang 2004] G. Fang, W. Gao et D. Zhao. *Large Vocabulary Sign Language Recognition Based on Fuzzy Decision Trees*. IEEE Transactions on Systems, Man, and Cybernetics - Part A : Systems and Humans, vol. 34, no. 3, pages 305–314, Mai 2004. (Cité en page 79.)

[Fusellier-Souza 2004] I. Fusellier-Souza. *Sémiogenèse des langues des signes : étude de langues des signes primaires (LSP) pratiquées par des sourds brésiliens*. PhD thesis, Université Paris 8, 2004. (Cité en pages 14, 15 et 16.)

[Futane 2012] Pravin R Futane et Rajiv V. Dharaskar. *Video gestures identification and recognition using Fourier descriptor and general fuzzy minmax neural network for subset of Indian sign language*. In 2012 12th International Conference on Hybrid Intelligent Systems (HIS), pages 525–530. IEEE, DÈcembre 2012. (Cité en page 79.)

[Gonzalez-Preciado 2012] M. Gonzalez-Preciado. *Computer vision methods for unconstrained gesture recognition in the context of sign language annotation*. 2012. (Cité en page 6.)

[Guillemot 1999] F Huet Guillemot. *Fusion d'images segmentées et interprétées. Application aux images aériennes*. PhD thesis, Cergy Pontoise, Toulouse, 1999. (Cité en page 82.)

[Hansen 2005] Dan Witzner Hansen et Arthur E.C. Pece. *Eye tracking in the wild*. Computer Vision and Image Understanding, vol. 98, no. 1, pages 155–181, Avril 2005. (Cité en page i.)

[Hansen 2010] Dan Witzner Hansen et Qiang Ji. *In the eye of the beholder : a survey of models for eyes and gaze*. IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 3, pages 478–500, Mars 2010. (Cité en page i.)

[Hieu 2008] Duong Van Hieu. *Vietnamese sign language recognition for hearing impaired peopl usong fuzzy hidden Markov models (FHMMs)*. PhD thesis, King Mongkut's university of technology north Bangkok, 2008. (Cité en page 79.)

[Holden 1996] EJ Holden, Robyn Owens et GG Roy. *Hand movement classification using an adaptive fuzzy expert system*, 1996. (Cité en page 82.)

[Holden 1999] Eun-jung Holden, Robyn Owens et GG Roy. *Adaptive fuzzy expert system for sign recognition*. Rapport technique, International Conference on Signal and Image Processing, Las Vegas, 1999. (Cité en page 79.)

[Jadon 2001] R.S. Jadon, Santanu Chaudhury et K.K. Biswas. *A fuzzy theoretic approach for video segmentation using syntactic features*. Pattern Recognition Letters, vol. 22, no. 13, pages 1359–1369, Novembre 2001. (Cité en page 82.)

[Kaindl 2006] Hermann Kaindl et J Falb. *From Requirements to Design : Model-driven Transformation or Mapping ?*, 2006. (Cité en page 83.)

[Lefebvre-Albaret 2010] F. Lefebvre-Albaret. *Traitement automatique de vidéos en LSF Modélisation et exploitation des contraintes phonologiques du mouvement*. PhD thesis, Université Paul Sabatier Toulouse III, 2010. (Cité en pages 6, 7, 83 et 84.)

[Lejeune 2004] F. Lejeune. *Analyse sémantico-sognitive d'énoncés en Langue des Signes Française pour une génération automatique de séquences getuelles*. PhD thesis, Paris 4, Paris, 2004. (Cité en page 11.)

[Leroy 2010] E. Leroy. *Didactique de la Langue des Signes Française : Attitudes et stratégies pédagogiques de l'enseignant sourd*. PhD thesis, Université Paris 8, 2010. (Cité en pages 11 et 16.)

[Lu 2011] Feng Lu, T Okabe, Y Sugano et Y Sato. *A head pose-free approach for appearance-based gaze estimation*. In Emanuele Hoey, Jesse and McKenna, Stephen and Trucco, editeur, Proceedings of the British Machine Vision Conference, pages 126.1–126.11. BMVA Press, 2011. (Cité en page ii.)

[MacLaughlin 1997] D MacLaughlin. *The structure of determiner phrases : Evidence from American Sign Language*. 1997. (Cité en page 16.)

[Matsuyama 1988] T. Matsuyama. *Expert systems for image processing-knowledge-based composition of image analysis processes*. In Pattern Recognition, 1988., 9th International Conference on, pages 125 –133 vol.1, nov 1988. (Cité en page 80.)

[Meurant 2003] L. Meurant. *L'anaphore syntaxique redéfinie au regard d'une langue des signes étude contrastive de structures anaphoriques.* La linguistique de la LSF : recherches actuelles French Sign Language Linguistics. Current Researchs, page 37, 2003. (Cité en pages 11 et 15.)

[Meurant 2005] Laurence Meurant. *De la deixis en langue des signes : le regard du locuteur*. Deixis : de l'énoncé à l'énonciation et vice-versa, pages 1–15, 2005. (Cité en page 17.)

[Meurant 2007] Laurence Meurant. *The Speaker's Eye Gaze Creating deictic, anaphoric and pseudo-deictic spaces of reference*. In Quadros, editeur, Theoretical Issues in Sign Language Research, pages 403–414, Florianopolis, Brazil., 2007. (Cité en page 15.)

[Nakamatsu 2012] Yukari Nakamatsu, Tetsuya Takiguchi et Yasuo Ariki. *Gaze Estimation Using 3D Active Appearance Models*. me.cs.scitec.kobe-u.ac.jp, pages 112–115, 2012. (Cité en page ii.)

[Padden 1983] C Padden. *Interaction of morphology and syntax in American Sign Language*. PhD thesis, University of California, 1983. (Cité en page 14.)

[Papazoglou 2010] Sophie Papazoglou, Parcours Recherche, Sophie Papazoglou Directeur et Annelies Braffort Organisme. *Fonctionnement du regard dans la syntaxe en Langue des Signes Française*. Master2r, Paris Ouest, Juillet 2010. (Cité en page 49.)

[Parisot 2003] Anne-Marie Parisot. *Explication unifiée de l'accord verbal en langue des signes québécoise : La notion de trace spatiale*. In Journée d'études internationales. La linguistique de la LSF : recherches actuelles, Lille, 2003. Université de Lille 3. (Cité en page 15.)

[Parisot 2006] Anne-Marie Parisot et Julie Rinfret. *La variation dans le marquage de la spécificité en langue des signes québécoise*. In Lille3, editeur, Colloque International Syntaxe, interprétation, lexique des langues signées, Lille, 2006. (Cité en page 16.)

[Parisot 2011] A.M. Parisot et Julie Rinfret. *Formes et fonctions des comportements du tronc et de la tête. Description comparée de production de trois signeurs (ASL, LSQ et LSF)*. In Séminaire international Multimodalité, description/analyse de langue des signes, évaluations, Grenoble, 2011. (Cité en page 29.)

[Phitakwinai 2008] Suwannee Phitakwinai, Sansanee Auephanwiriyakul et Nipon Theera-Umpon. *Thai sign language translation using fuzzy c-means and scale invariant feature transform*. In International Conference on Computational Science and Its Applications, pages 1107–1119, Berlin Heidelberg, 2008. LNCS 5073. (Cité en page 79.)

[Pit 2004] *Expert knowledge-guided segmentation system for brain {MRI}*. NeuroImage, vol. 23 Supplement 1, no. 0, pages S85 – S96, 2004. (Cité en page 80.)

[Radice 1985] R.A Radice, N.K Roth, A.C.Jr O'Hara et W.A Ciarfella. *A Programming Process Architecture*. IBM Systems Journal, vol. 24, no. 2, pages 79–90, 1985. (Cité en page 98.)

[Rijsbergen 1979] C. J. Van Rijsbergen. Information retrieval. Butterworth-Heinemann, Newton, MA, USA, 2nd édition, 1979. (Cité en page 64.)

[Rinfret 2009] J. Rinfret. *L'association du nom en langue des signes québécoises : Formes, fonctions et sens*. PhD thesis, UQAM, 2009. (Cité en pages 11, 14 et 16.)

[Risler 2005] A. Risler. *Construction/déconstruction de l'espace de signation*. TALN, 2005. (Cité en pages 11, 16 et 17.)

[Salam 2012] H Salam. *A Multi-texture approach for estimating iris positions in the eye using 2.5 D active appearance model*. In IEEE, editeur, Proceedings of the IEEE 2012 International Conference on Image Processing, Orlando, 2012. (Cité en page ii.)

[Sarfraz 2005] M Sarfraz, Yusuf A Syed et M Zeeshan. *A System for Sign Language Recognition Using Fuzzy Object Similarity Tracking*. 2010 14th International Conference Information Visualisation, vol. 0, pages 233–238, 2005. (Cité en pages 79 et 82.)

[Shih 2000] SW Shih, YT Wu et Jin Liu. *A calibration-free gaze tracking technique*. Pattern Recognition, pages 201–204, 2000. (Cité en page i.)

[Shih 2004] Sheng-Wen Shih et Jin Liu. *A novel approach to 3-D gaze tracking using stereo cameras.* IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, vol. 34, no. 1, pages 234–45, FÈvrier 2004. (Cité en page ii.)

[Takatani 2010] M Takatani, Yasuo Ariki et Tetsuya Takiguchi. *Gaze Estimation Using Regression Analysis and AAMs Parameters Selected Based on Information Criterion*. In Gaze Sensing and Interactions, 2010. (Cité en page ii.)

[Tomasi 1992] Carlo Tomasi et T Kanade. *Shape and motion from image streams : a factorization method : full report on the orthographic case*. International Journal of Computer Vision, vol. 9, no. 7597, pages 137–154, 1992. (Cité en page 84.)

[Valenti 2010] Roberto Valenti, Adel Lablack, Nicu Sebe, Chabane Djeraba et Theo Gevers. *Visual Gaze Estimation by Joint Head and Eye Information*. 2010 20th International Conference on Pattern Recognition, pages 3870–3873, Ao˚t 2010. (Cité en page i.)

[Valenti 2012] Roberto Valenti, Nicu Sebe et Theo Gevers. *Combining head pose and eye location information for gaze estimation.* IEEE transactions on image processing : a publication of the IEEE Signal Processing Society, vol. 21, no. 2, pages 802–15, FÈvrier 2012. (Cité en page ii.)

[Viel 2003] E. Viel. La marche humaine, la course et le saut : biomécanique, explorations, normes et dysfonctionnements. Le Point en rééducation. Masson, 2003. (Cité en page 26.)

[Wallhoff 2006] Frank Wallhoff, Markus AblaBmeier et Gerhard Rigoll. *Multimodal Face Detection, Head Orientation and Eye Gaze Tracking*. In 2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, pages 13–18. IEEE, Septembre 2006. (Cité en page ii.)

[Winston 1995] EA Winston. *Spatial mapping in comparative discourse frames*. Language, gesture, and space, pages 1–21, 1995. (Cité en page 16.)

[Xu 2009] Guoqing Xu, Yangsheng Wang, Jituo Li et Xiaoxu Zhou. *Real Time Detection of Eye Corners and Iris Center from Images Acquired by Usual Camera*. In 2009 Second International Conference on Intelligent Networks and Intelligent Systems, pages 401–404. Second International Conference on Intelligent Networks and Intelligent Systems, IEEE, Novembre 2009. (Cité en page i.)

Résumé : Cette thèse porte sur le rôle et l'analyse du regard en langue des signes où celui-ci joue un rôle important. Dans toute langue, le regard permet de maintenir la relation de communication. En langue des signes, il permet, en plus, de structurer le discours ou l'interaction entre locuteurs, en s'investissant dans des fonctions linguistiques complexes.

Nous nous intéressons au rôle de référencement qui consiste à mettre le focus sur un élément du discours. En langue des signes, les éléments du discours sont spatialisés dans l'espace de signation ; ainsi, mettre le focus sur un élément du discours revient à identifier et activer son emplacement spatial (locus), ce qui va mobiliser un ou plusieurs composants corporels, les mains, les épaules, la tête et le regard. Nous avons donc analysé le concept de référencement sous ses formes manuelles et / ou non manuelles et avons mis en place un système de reconnaissance de structures de référencement qui prend en entrée une vidéo en langue des signes. Le système de reconnaissance consiste en trois étapes : 1) la modélisation 3D du concept de référencement, 2) la transformation du modèle 3D en un modèle d'aspect exploitable par un programme de traitement 2D et 3) la détection, qui utilise ce modèle d'aspect.

La modélisation consiste en l'extraction de caractéristiques gestuelles du concept de référencement à partir de corpus composés de capture 3D de mouvement et du regard et annotés manuellement à partir de vidéos. La modélisation concerne la description des composantes corporelles qui jouent un rôle dans le référencement et la quantification de quelques propriétés gestuelles des composantes corporelles en question. Les modèles obtenus décrivent : 1) La dynamique du mouvement de la main dominante et 2) la proximité spatiale entre des composantes corporelles et l'élément discursif spatialisé.

La mise en œuvre de la méthode de reconnaissance intègre ces modèles 3D de profil dynamique de la main et de variation de distance entre composantes corporelles et l'élément discursif ainsi que le modèle temporel de décalages entre mouvements. Etant donné que les modèles obtenus sont tridimensionnels et que l'entrée du système de reconnaissance de structures de référencement est une vidéo 2D, nous proposons une transformation des modèles 3D en 2D afin de permettre leur exploitation dans l'analyse de la vidéo 2D et la reconnaissance des structures de référencement. Nous pouvons alors appliquer un algorithme de reconnaissance à ces corpus vidéo 2D. Les résultats de reconnaissance sont sous la forme d'intervalles temporels. On constate la présence de deux variantes principales de référencement. Ce travail pionnier sur la caractérisation et la détection des référencements nécessiterait d'être approfondi sur des corpus beaucoup plus importants, cohérents et riches et avec des méthodes plus élaborées de classification. Cependant il a permis d'élaborer une méthodologie d'analyse réutilisable.

Mots clés : Corpus en langue des signes, Marquage spatial, Modélisation gestuelle, Segmentation d'images, Apprentissage automatique, Reconnaissance.

Abstract : This thesis focuses on the role and analysis of gaze in sign language where it plays an important role. In any language, the gaze keeps the communication relationship. In addition to that, it allows structuring a sign language discourse or interaction between signers, by investing in complex linguistic features.

We focus on the role of reference, which is to put the focus on an element of the discourse. In sign language, the components of the discourse are localized in the signing space ; thus putting the focus on an element of discourse which is to identify and activate its spatial location (locus), which will mobilize one or more body parts, hands, shoulders, head and eyes. We therefore analyzed the concept of reference in its manual and / or non-manual gestures and set up a reference-based recognition system that takes as input a video in sign language. The recognition system consists of three steps : - 3D modeling of the concept of reference. - The transformation of the 3D model into a 2D model useable by a 2D recognition system. - The detection system, which uses this 2D model.

Modeling involves the extraction of gestural characteristics of the concept of reference from corpus consisted of 3D motion capture and gaze and manually annotated videos and the temporal pattern of time lags between motions. Modeling concerns the description of body parts that play a role in reference and the quantification of their gestural. The resulting models describe : 1) The dynamic movement of the dominant hand and 2) the distances between body parts and locus and 3) the time lags between the beginning of motions.

The implementation of the recognition method integrates these 3D models. Since the resulting models are three-dimensional and the recognition system has, as input, a 2D video, we propose a transformation of 3D models to 2D to allow their use in the analysis of 2D video and in pattern recognition of reference structures.

We can then apply a recognition algorithm to the 2D video corpus. The recognition results are a set of time slots with two main variants of reference.

This pioneering work on the characterization and detection of references structures would need to be applied on much larger corpus, consistent and rich and more sophisticated classification methods. However, it allowed to make a reusable methodology of analysis.

Keywords : Sign Language corpus, Spatial markers, Gesture modeling, Image Segmentation, Machine learning, Recognition